(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平11－212978

(43)公開日　平成11年(1999)８月６日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ

審査請求　未請求　請求項の数23　ＯＬ　（全 20 頁）

(21)出願番号　特願平10－14314

(22)出願日　平成10年(1998)１月27日

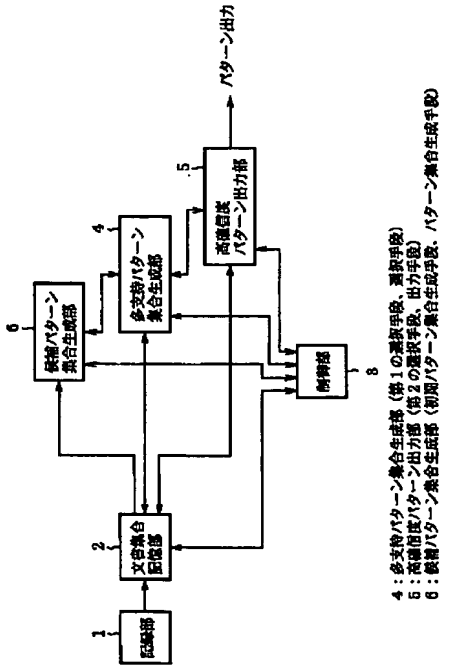
(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号

(72)発明者　小中　裕喜
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

(74)代理人　弁理士　田澤　博昭　（外１名）

(54)【発明の名称】　文書情報解析方法、文書情報解析装置および記録媒体

(57)【要約】
【課題】　他のクラスタと比較してクラスタ特有のパターンを抽出する。
【解決手段】　制御部３により指定されたクラスタに対応する各パターンについて、多支持パターン集合生成部４は、そのクラスタにおいてパターンが出現する文書情報の数とクラスタに属する全ての文書情報の数との比である支持率を計算し、その支持率が所定の最小支持率以上のパターンを選択し、高確信度パターン出力部５に出力する。高確信度パターン出力部５は、そのクラスタにおいてパターンが出現する文書情報の数と全てのクラスタにおいてパターンが出現する文書情報の数との比である確信度を計算し、確信度が所定の最小確信度以上であるパターンを選択する。



1

【特許請求の範囲】

【請求項１】　複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成するステップと、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択するステップと、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択するステップと、

前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力するステップと、

前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するステップとを備えた文書情報解析方法。

【請求項２】　所定のパターンの確信度の計算において、前記所定のパターンが出現した文書情報の文書ＩＤを記憶し、

所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の文書ＩＤを読み出し、その文書情報においてだけ、前記（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントすることを特徴とする請求項１記載の文書情報解析方法。

【請求項３】　所定のパターンの確信度の計算において、前記所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、

所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、前記（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントすることを特徴とする請求項１記載の文書情報解析方法。

2

【請求項４】　初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上となるキーワード以外のキーワードを、すべての文書情報から除去することを特徴とする請求項１から請求項３のうちのいずれか１項記載の文書情報解析方法。

【請求項５】　各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を前記元の文書情報の代わりに使用することを特徴とする請求項１から請求項４のうちのいずれか１項記載の文書情報解析方法。

【請求項６】　所定のパターン集合は、初期のパターン集合であることを特徴とする請求項５記載の文書情報解析方法。

【請求項７】　新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去することを特徴とする請求項１から請求項６のうちのいずれか１項記載の文書情報解析方法。

【請求項８】　初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが前記高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する前記高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、前記高確信度パターン集合に追加することを特徴とする請求項１から請求項７のうちのいずれか１項記載の文書情報解析方法。

【請求項９】　多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、

前記確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力することを特徴とする請求項１から請求項７のうちのいずれか１項記載の文書情報解析方法。

【請求項１０】　クラスタ毎に一括して、所定の順番でパターンを出力することを特徴とする請求項８または請求項９記載の文書情報解析方法。

【請求項１１】　複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキ

ーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成するステップと、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加するステップと、

前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するステップと、

前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するステップとを備えた文書情報解析方法。

【請求項１２】　初期の多支持パターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上となるキーワード以外のキーワードを、すべての文書情報から除去することを特徴とする請求項１１記載の文書情報解析方法。

【請求項１３】　累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を前記元の文書情報の代わりに使用することを特徴とする請求項１１または請求項１２記載の文書情報解析方法。

【請求項１４】　クラスタ毎に一括して、所定の順番で、確信度が所定の最小確信度以上である多支持パターンを出力することを特徴とする請求項１１から請求項１３のうちのいずれか１項記載の文書情報解析方法。

【請求項１５】　所定の順番は、支持率の値に応じた順番であることを特徴とする請求項１０または請求項１４記載の文書情報解析方法。

【請求項１６】　所定の順番は、確信度の値に応じた順番であることを特徴とする請求項１０または請求項１４記載の文書情報解析方法。

【請求項１７】　所定の順番は、パターンを構成するキーワードの数に応じた順番であることを特徴とする請求項１０または請求項１４記載の文書情報解析方法。

【請求項１８】　所定の順番は、各パターンの条件付きエントロピーの値に応じた順番であることを特徴とする請求項１０または請求項１４記載の文書情報解析方法。

【請求項１９】　すべてのクラスタから、パターンを出力するクラスタを選択するステップを備え、

選択したクラスタだけについてパターンを選択し、出力することを特徴とする請求項１から請求項１８のうちのいずれか１項記載の文書情報解析方法。

【請求項２０】　複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第１の選択手段と、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第２の選択手段と、

前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力する出力手段と、

前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段とを備えた文書情報解析装置。

【請求項２１】　複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ

１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段と、

前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段と、

前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段とを備えた文書情報解析装置。

【請求項２２】　コンピュータを、

複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第１の選択手段、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第２の選択手段、

前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力する出力手段、

前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段として機能させるためのプログラムを記録した記録媒体。

【請求項２３】　コンピュータを、

複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段、

前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段、

前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段として機能させるためのプログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】
【発明の属する技術分野】この発明は、複数のクラスタに分類して記録されている、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報において、各クラスタについて、そのクラスタを特徴づけるキーワードの連言（パターン）を抽出するための文書情報解析方法、文書情報解析装置および記録媒体に関するものである。
【０００２】
【従来の技術】近年の情報通信メディア、特にインターネットやWWW（World Wide Web）などの発展に伴い、容易に入手可能な情報の量が急激に増加している。例えば、WWWにおいて各種情報を検索するために、Ｙａｈｏｏ（商標）などの検索エンジンが提供されている。
【０００３】従来、文書データベースにおいて検索エンジンなどを利用して所望の文書情報を検索する場合、１つまたは複数のキーワードを指定し、そのキーワードに該当する文書情報を文書データベースから読み出すようにしている。
【０００４】しかしながら、大量の種々雑多な文書情報がそのまま記録されている場合、その文書情報の中から所望の文書情報を抽出するために、すべての文書情報を検査する必要があり実用的ではない。例えば、検索対象に対する利用者の知識が曖昧であり、利用者がその対象を的確にとらえるキーワードを想起することが困難である場合、利用者は抽象的なキーワードを使用して検索するため、関連性の低い長大な文書情報が抽出されることが多く、利用者にとって興味のある文書情報に到達するまでに長い時間を費やしてしまうことになる。
【０００５】そこで、オペレータが予めそのような種々雑多な文書情報をその内容に基づいて分類し、内容に関連性があるもの同士をクラスタごとにまとめて記録することや、ＳＯＭ（Self-Organizing Maps）などの自己組織化アルゴリズムを使用して、種々雑多な文書情報をクラスタに分類することが考えられる。なお、自己組織化アルゴリズムには、例えば「Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration」（Ｌａｇｕｓら著、「Proc.of the Second International Conference on Knowledge Discovery and Data Mining KDD'96」の第２３８頁～第２４３頁、１９９６年）に記載のものや、「Exploration of Document Collections with Self-Organizing Maps: A Novel Approach to Similarity Representation」（Ｍｅｒｋｌ著、「Proc. of the firstEuropean Symposium on Principles of Data Mining and Knowledge DiscoveryPKDD'97」の第１０１頁～第１１１頁、１９９７年）に記載のものがある。
【０００６】このように、文書情報をクラスタに分類しておくことにより、所望の文書情報を検索する場合、まず、その文書情報に関連するクラスタを検索し、該当するクラスタだけを検査すればよいことになる。
【０００７】しかしながら、文書情報をその詳細な内容にわたってクラスタリングする場合、オペレータによるクラスタリングでは、時間やコストが多く必要になることになる。一方、ＳＯＭなどの自己組織化アルゴリズムを使用したクラスタリングでは、各クラスタがどのような特徴に基づいて生成されたのか、また各クラスタの文書情報がどのような理由でそのクラスタに分類されたのかを判断することが困難な場合が多い。
【０００８】近年、データベースから特徴的なパターンを抽出するいわゆるデータマイニング技術が提案されている。このデータマイニング技術により、所定の文書情報から、その文書情報において特徴的なキーワードのパターンを抽出することができる。したがって、特徴的なキーワードのパターンを抽出しておき、その文書情報に関連づけて記録しておくことにより、そのパターンに基づいて文書情報の検索を行えば、比較的容易に所望の文書情報に到達することになると考えられる。
【０００９】そのようなデータマイニングの方法としては、例えば、「Fast Algorithms for Mining Association Rules」（Ａｇｒａｗａｌら著、「Proc. of VLDB'94」の第４８７頁～第４９９頁、１９９４年）（以下、文献１という）に記載のものや「An Effective Hash Based Algorithm for Mining Association Rules」（Ｐａｒｋら著、「Proc. of SIGMOD'95」の第１７５頁～第１８６頁、１９９５年）（以下、文献２という）に記載のものがある。さらに文書情報に適用したものとしては、「Pattern Based Browsing in Document Collections」（Ｆｅｌｄｍａｎら著、「Proc. of the First European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97」の第１１２頁～第１２２頁、１９９７年）（以下、文献３という）に記載のものがある。
【００１０】
【発明が解決しようとする課題】文書情報から特徴的なキーワードのパターンを抽出する従来の文書解析方法は以上のように構成されているので、例えば文献３に記載の方法により、個々のクラスタ内において特徴的なパターンを抽出することはできるが、他のクラスタのパターンと比較して特徴的なそのクラスタ特有のパターンを抽出することが困難であるなどの課題があった。
【００１１】すなわち、複数の類似するクラスタが存在する場合に、文書情報の検索を容易にするようなそのクラスタ特有のパターンを抽出することが困難であるという課題があった。
【００１２】例えば、図９は、スポーツ関係の新聞記事を記録したデータベースの一例を示している。図９に示すデータベースにおいては、サッカー関連の文書情報と、高校総体関連の文書情報が、各クラスタに対応するファイルに記録されている。この場合、例えば「つり

輪」という単一のキーワードにより構成されるパターンは、「高校総体総合」クラスタと「高校総体つり輪」クラスタにおいて多く出現するので、このパターンにより、「高校総体総合」クラスタまたは「高校総体つり輪」クラスタを特徴づけることは困難である。一方、例えば「つり輪」と「種目別」により構成されるパターンは、「高校総体つり輪」クラスタにおいて多く出現するので、このパターンにより、「高校総体つり輪」クラスタを特徴づけることが可能である。しかしながら、このような、クラスタ特有のパターン、例えば上述の「つり輪」と「種目別」とにより構成されるパターンを抽出することは、個々のクラスタ内だけにおいて特徴的なパターンを抽出する従来の方法では困難であった。

【００１３】この発明は上記のような課題を解決するためになされたもので、複数のクラスタに分類して記録されている、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報において、各クラスタについて、そのクラスタを特徴づけるキーワードのパターンを抽出するための文書情報解析方法、文書情報解析装置および記録媒体を得ることを目的とする。

【００１４】

【課題を解決するための手段】この発明に係る文書情報解析方法は、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択し、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力し、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するものである。

【００１５】この発明に係る文書情報解析方法は、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の文書ＩＤを記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の文書ＩＤを読み出し、その文書情報においてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするものである。

【００１６】この発明に係る文書情報解析方法は、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするものである。

【００１７】この発明に係る文書情報解析方法は、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを、すべての文書情報から除去するものである。

【００１８】この発明に係る文書情報解析方法は、各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【００１９】この発明に係る文書情報解析方法は、各クラスタについて、初期のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【００２０】この発明に係る文書情報解析方法は、新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去するものである。

【００２１】この発明に係る文書情報解析方法は、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけ

を、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加するものである。

【００２２】この発明に係る文書情報解析方法は、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、その確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力するものである。

【００２３】この発明に係る文書情報解析方法は、クラスタ毎に一括して、所定の順番でパターンを出力するものである。

【００２４】この発明に係る文書情報解析方法は、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加し、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するものである。

【００２５】この発明に係る文書情報解析方法は、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードをすべての文書情報から除去するものである。

【００２６】この発明に係る文書情報解析方法は、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【００２７】この発明に係る文書情報解析方法は、クラスタ毎に一括して、所定の順番で、確信度が所定の最小確信度以上であるパターンを出力するものである。

【００２８】この発明に係る文書情報解析方法は、支持率の値に応じた順番でパターンを出力するものである。

【００２９】この発明に係る文書情報解析方法は、確信度の値に応じた順番でパターンを出力するものである。

【００３０】この発明に係る文書情報解析方法は、パターンを構成するキーワードの数に応じた順番でパターンを出力するものである。

【００３１】この発明に係る文書情報解析方法は、各パターンの条件付きエントロピーの値に応じた順番でパターンを出力するものである。

【００３２】この発明に係る文書情報解析方法は、すべてのクラスタから選択されたクラスタだけについてパターンを選択し、出力するものである。

【００３３】この発明に係る文書情報解析装置は、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第１の選択手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第２の選択手段と、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力する出力手段と、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段とを備えたものであ

る。

【００３４】この発明に係る文書情報解析装置は、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段と、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段と、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段とを備えたものである。

【００３５】この発明に係る記録媒体は、コンピュータを、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第１の選択手段、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第２の選択手段、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力する出力手段、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段として機能させるためのプログラムを記録したものである。

【００３６】この発明に係る記録媒体は、コンピュータを、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段として機能させるためのプログラムを記録したものである。

【００３７】

【発明の実施の形態】以下、この発明の実施の一形態を

説明する。

実施の形態１．図１は、この発明の実施の形態１による文書情報解析装置の構成を示すブロック図である。図において、１は、例えば図９に示すような文書情報をクラスタに分類して、各クラスタに対応するファイルにそれぞれ対応する文書情報を記録する例えばハードディスク装置、フロッピーディスクを装着したフロッピーディスク駆動装置などの記録部である。なお、文書情報は、図９に示すように文書ＩＤと、１つまたは複数のキーワードで構成される。

【００３８】２は、記録部１に記録された文書情報をクラスタ毎に読み出して記憶するとともに、制御部３により指定されたクラスタに対応する文書情報を候補パターン生成部６に出力する文書集合記憶部である。また文書集合記憶部２は、多支持パターン集合生成部４または高確信度パターン出力部５によりクラスタとパターンが指定されると、指定されたクラスタにおいて、指定されたパターンが出現する文書情報の数を計算し、多支持パターン集合生成部４または高確信度パターン出力部５に出力する。なお、文書集合記憶部２は、各文書情報において、指定されたパターンが出現するか否かを検査する場合、その文書情報を構成するキーワードの部分集合がそのパターンに一致するか否かを検査する。したがって、この検査を効率的に行うために、ハッシュテーブルやハッシュ木を用いて各文書のキーワード集合に対して部分集合となる可能性のあるパターンを絞り込んだり、各パターン、文書情報においてキーワード集合を整列したり、あるいはビットパターンで表すようにしてもよい。なお、文書集合記憶部２を所定のコンピュータネットワークに接続し、そのコンピュータネットワークを介して他の記録部から文書情報を読み出すようにしてもよい。

【００３９】６は、文書集合記憶部２より供給されたあるクラスタの文書情報に用いられている各キーワードに対し、１つのキーワードにより構成される候補パターンを生成して、多支持パターン集合生成部４に出力するとともに、多支持パターン集合生成部４より供給され、その多支持パターン集合の多支持パターンを構成するキーワードの数をｎとしたとき、（ｎ－１）個のキーワードが共通する任意の２つの多支持パターンから（ｎ＋１）個のキーワードからなるパターンを生成し、そのパターンの任意のｎ個のキーワードが多支持パターン集合に含まれるものを候補パターンとして、新たな候補パターン集合を生成し、多支持パターン集合生成部４に出力する候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）である。

【００４０】４は、候補パターン集合生成部６より供給される各候補パターンについて、制御部３により指定されたクラスタにおいてその候補パターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を、文書集合記憶部２を利用して計算し、その支持率が所定の最小支持率以上であるパターンを新たな多支持パターンとして選択し、それらの多支持パターンの集合を高確信度パターン出力部５と候補パターン集合生成部６に出力する多支持パターン集合生成部（第１の選択手段、選択手段）である。

【００４１】５は、多支持パターン集合生成部４より供給される各パターンについて、制御部３により指定されたクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を、文書集合記憶部２を利用して計算し、多支持パターン集合生成部４より供給されたパターン集合から、確信度が所定の最小確信度以上であるパターンを選択し、そのパターンを図示せぬ装置に出力する高確信度パターン出力部（第２の選択手段、出力手段）である。なお、高確信度パターン出力部５は、図示せぬ装置に出力するかわりに、そのパターンを、指定されたクラスタに関連づけて記録部１に記録させるようにしてもよい。

【００４２】３は、文書集合記憶部２、多支持パターン集合生成部４、高確信度パターン出力部５、および候補パターン集合生成部６を制御する制御部である。

【００４３】なお、上述の文書集合記憶部２、制御部３、多支持パターン集合生成部４、高確信度パターン出力部５、および候補パターン集合生成部６は、各種電子回路でハードウェアとして構成してもよいが、ＣＰＵを文書集合記憶部２、制御部３、多支持パターン集合生成部４、高確信度パターン出力部５、および候補パターン集合生成部６として機能させるためのプログラムを記録した記録媒体（例えば、ハードディスク装置、フロッピーディスクなどの磁気ディスク、コンパクトディスクなどの光ディスク）から、そのプログラムを読み取ることが可能なコンピュータで構成してもよい。

【００４４】次に動作について説明する。図２は、図１の文書情報解析装置の動作を説明するフローチャートである。まず、ステップＳＴ１において、文書集合記憶部２は、記録部１に記録された文書情報を読み出し、候補パターン集合生成部６に出力する。

【００４５】ステップＳＴ２において、制御部３は、文書集合記憶部２により文書情報を読み出されたクラスタのうち、後述のステップＳＴ３～ステップＳＴ１０の処理を実行されていないクラスタがあるか否かを判定し、すべてのクラスタについてその処理を実行したと判定した場合、パターン抽出の処理を終了し、その処理を実行されていないクラスタが少なくとも１つあると判定した場合、そのクラスタのうちの１つを指定する。

【００４６】候補パターン集合生成部６は、ステップＳＴ３において、供給された文書情報に含まれる各キーワードを１つのキーワードからなる候補パターン（初期候補パターン）とし、クラスタ毎に候補パターン集合（初期候補パターン集合）を生成し、多支持パターン集合生

成部４に出力する。そして、ステップＳＴ４において、多支持パターン集合生成部４は、制御部３により指定されたクラスタに属する各候補パターンについて、文書集合記憶部２を制御してそのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップＳＴ５において、そのクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率をそれぞれ計算し、その支持率が所定の最小支持率以上であるパターンを多支持パターンとして選択し、それらの多支持パターンの集合を多支持パターン集合として高確信度パターン出力部５と候補パターン生成部６に出力する。

【００４７】次に、高確信度パターン出力部５は、ステップＳＴ６において、多支持パターン集合生成部４より供給された多支持パターンについて、文書集合記憶部２を制御して、指定されたクラスタ以外の残りのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップＳＴ７において、そのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、その確信度が所定の最小確信度以上であるパターンを選択し、図示せぬ装置に出力する。すなわち、指定されたクラスタにおいてそのパターンが出現する文書情報の数と、すべてのクラスタにおいてそのパターンが出現する文書情報の数とに基づいた確信度を利用することにより、そのクラスタ特有のパターンが選択される。

【００４８】そして、ステップＳＴ８において、制御部３は、多支持パターン集合生成部４により選択された多支持パターンの数が所定の最小パターン数より小さいか否かを判定し、選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、指定したクラスタに対する処理を終了し、ステップＳＴ２に戻る。一方、選択されたパターンの数が所定の最小パターン数以上であると判定した場合、制御部３は、ステップＳＴ９において、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したか否かを判定し、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、指定したクラスタに対する処理を終了し、ステップＳＴ２に戻る。

【００４９】一方、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達していないと判定した場合、制御部３は、候補パターン集合生成部６に制御信号を供給する。ステップＳＴ１０において、候補パターン集合生成部６は、その制御信号を受信すると、多支持パターン集合生成部４より供給された多支持パターン集合について、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターン（（ｎ－１）個のキーワードが共通である２つの多支持パターン）を構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、そのパターン集合を候補パターン集合として多支持パターン集合生成部４に出力する。なお、このステップＳＴ１０の処理には、文献１に記載されている「apriori-gen 」関数を利用することができる。

【００５０】そして、ステップＳＴ４に戻り、ステップＳＴ１０で生成された候補パターン集合について、ステップＳＴ４～ステップＳＴ１０の処理が実行される。

【００５１】すなわち、ステップＳＴ５～ステップＳＴ１０の処理を反復する毎に、パターンを構成するキーワードの数を１つずつ増えていくとともに、それに伴い、支持率が所定の最小支持率以上であるパターンが次第に減少していく。

【００５２】以上のように、この実施の形態１によれば、各クラスタについて、所定のパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率と、そのクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度とに基づいて、キーワードのパターンを抽出するので、そのクラスタ特有のパターンを抽出することができるという効果が得られる。

【００５３】なお、ステップＳＴ４～ステップＳＴ１０の反復処理において、２回目以降の処理では、前回の処理において選択された多支持パターンを有していた文書情報またはクラスタだけについて、候補パターンが出現するか否かの検査を行うようにしてもよい。前回の処理において選択された多支持パターンを有しない文書情報は、そのパターンを含む今回の処理の候補パターンも同様に有しないので、このようにすることにより、処理する文書情報やクラスタの数が低減され、処理時間を短縮することができるという効果が得られる。

【００５４】なお、オペレータが図示せぬ操作部を操作して選択したクラスタだけについて、パターンの抽出の処理を行うようにしてもよい。その場合、図示せぬ操作部は、制御部３に接続され、その操作に対応する信号に応じて制御部３は、各部を制御する。このようにすることにより、不要なクラスタに対して処理が行われないので、処理時間を短縮することができるという効果が得られる。

【００５５】実施の形態２．この発明の実施の形態２による文書情報解析装置は、実施の形態１による文書情報解析装置の文書集合記憶部２と多支持パターン集合生成部４に後述の機能を追加したものである。

【００５６】すなわち、実施の形態２における文書集合記憶部２は、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情

報から除去する。また、実施の形態２における多支持パターン集合生成部４は、候補パターン集合生成部６より供給される候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【００５７】なお、その他の構成要素は、実施の形態１のものと同様であるので、その説明を省略する。

【００５８】次に動作について説明する。図３は、実施の形態２による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０の処理は、実施の形態１のものと同様であるので、その説明を省略する。

【００５９】実施の形態２においては、ステップＳＴ１の後のステップＳＴ２１において、文書集合記憶部２は、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。

【００６０】また、ステップＳＴ５の後のステップＳＴ２２において、多支持パターン集合生成部４は、候補パターン集合生成部６より供給された候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【００６１】なお、ステップＳＴ２２において更新された文書情報は、各クラスタについての処理が終了したときに消去され、次のクラスタについての処理は、ステップＳＴ２１において生成された文書情報に基づいて行われる。

【００６２】以上のように、この実施の形態２によれば、パターンを構成する各キーワードについて支持率を計算し、その支持率が所定の最小支持率より低いキーワードを文書情報から適宜除去するようにしたので、パターンが出現するか否かを検査する対象になる文書情報の量を低減させることができ、処理時間を短縮することができるという効果が得られる。

【００６３】なお、ステップＳＴ２２の処理は、ステップＳＴ２２、ステップＳＴ４～ステップＳＴ１０の反復処理の最初の処理においてだけ行うようにしてもよい。２回目以降のステップＳＴ２２の処理においては、除去されるキーワードの数が少ないので、このようにすることにより、より処理時間を短縮することができる場合が多い。

【００６４】実施の形態３．この発明の実施の形態３による文書情報解析装置は、実施の形態２による文書情報解析装置の高確信度パターン出力部５に後述の機能を追加したものである。

【００６５】すなわち、実施の形態３における高確信度パターン出力部５は、多支持パターン集合生成部４より供給されたパターンから、その確信度が所定の最小確信度以上であるパターンを選択し、そのパターンを図示せぬ装置に出力した後、確信度が所定の最大確信度以上であるパターンを除去し、残りのパターンを新たに多支持パターン集合として多支持パターン集合生成部４に戻す。なお、多支持パターン集合生成部４は、高確信度パターン出力部５から戻された多支持パターン集合を候補パターン集合生成部６に出力する。

【００６６】なお、その他の構成要素は、実施の形態２のものと同様であるので、その説明を省略する。

【００６７】次に動作について説明する。図４は、実施の形態３による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０、ステップＳＴ２１、およびステップＳＴ２２の処理は、実施の形態２のものと同様であるので、その説明を省略する。

【００６８】実施の形態３においては、高確信度パターン出力部５は、ステップＳＴ７において、多支持パターン集合生成部４より供給されたパターンから、その確信度が所定の最小確信度以上であるパターンを選択し、図示せぬ装置に出力した後、ステップＳＴ３１において、その確信度が所定の最大確信度以上であるパターンを除去し、残りのパターンを新たに多支持パターン集合として多支持パターン集合生成部４に戻す。なお、多支持パターン集合生成部４は、ステップＳＴ６で多支持パターン集合を候補パターン集合生成部６に出力する代わりに、ステップＳＴ３１で高確信度パターン出力部５より戻された多支持パターン集合を出力する。

【００６９】以上のように、この実施の形態３によれば、確信度が所定の最大確信度以上であるパターンを多支持パターン集合から除去するようにしたので、そのようなパターンを含む、利用者にとって不必要に複雑なパターンの出力と、そのための処理を抑制することができるという効果が得られる。

【００７０】実施の形態４．この発明の実施の形態４による文書情報解析装置は、実施の形態３による文書情報解析装置の高確信度パターン出力部５に後述の機能を追加したものである。

【００７１】実施の形態４においては、高確信度パターン出力部５は、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、確信度がそのパターンのサブセットに該当する高確信度パターン集合におけるパターンの確信度以上であるパターンだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加する。

【００７２】なお、その他の構成要素は、実施の形態３のものと同様であるので、その説明を省略する。

【００７３】次に動作について説明する。図５は、実施の形態４による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０、ステップＳＴ２１、ステップＳＴ２２、およびステップＳＴ３１の処理は、実施の形態３のものと同様であるので、その説明を省略する。

【００７４】実施の形態４においては、高確信度パターン出力部５は、ステップＳＴ４１において、初期状態の空集合である高確信度パターン集合を記憶しておき、ステップＳＴ４２において、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力し、その後、高確信度パターン集合に追加する。

【００７５】以上のように、この実施の形態４によれば、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを選択するようにしたので、キーワード数が多くかつ確信度の低いパターンを除去することができ、キーワード数が少なく確信度の高いパターンが選択され、検索時の効率を向上させることができるという効果が得られる。

【００７６】なお、ステップＳＴ１０において新たなパターン集合を生成するときに、各パターンのすべてのサブセットの確信度を同時に計算し、そのパターンのすべてのサブセットの確信度のうちの最大値をそのパターンに対応して記憶して、ステップＳＴ４２においては、そのサブセットの確信度の最大値以上の確信度を有するパターンだけを選択し、そのパターンを出力するとともに、高確信度パターン集合に追加するようにしてもよい。このようにすることにより、ステップＳＴ４２における処理を簡略化することができ、処理時間を短縮することができるという効果が得られる。

【００７７】実施の形態５．この発明の実施の形態５による文書情報解析装置は、実施の形態４による文書情報解析装置の高確信度パターン出力部５に後述の機能を追加したものである。

【００７８】すなわち、実施の形態５においては、高確信度パターン出力部５は、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、そのパターンの確信度がそのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上であるパターンだけ高確信度パターン集合に追加していき、各クラスタに対する処理が終了したとき、クラスタ毎に一括して、所定の順番で高確信度パターン集合を出力する。

【００７９】なお、その他の構成要素は、実施の形態４のものと同様であるので、その説明を省略する。

【００８０】次に動作について説明する。図６は、実施の形態５による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０、ステップＳＴ２１、ステップＳＴ２２、ステップＳＴ３１、ステップＳＴ４１、およびステップＳＴ４２の処理は、実施の形態４のものと同様であるので、その説明を省略する。

【００８１】実施の形態５においては、高確信度パターン出力部５は、ステップＳＴ４２において、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、そのパターンの確信度がそのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上であるパターンだけ高確信度パターン集合に追加していく。なお、このときには、高確信度パターン出力部５は、そのパターンを図示せぬ所定の装置には出力しない。

【００８２】そして、高確信度パターン出力部５は、ステップＳＴ８またはステップＳＴ９において、選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、または、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、ステップＳＴ５１において、そのクラスタに対応するパターンとして、所定の順番で高確信度パターン集合を出力する。

【００８３】なお、このときの順番は、高確信度パターン集合を構成するパターンの支持率、確信度、パターンを構成するキーワードの数などに基づいて予め設定する。

【００８４】また、Ｄを文書情報の集合、Ｄｐをパターンｐが出現する文書情報の集合としたときの、式（１）および式（２）により規定される各パターンｐに対する所定のクラスタc_iの条件付きエントロピーEnt_i（Ｄ｜ｐ）の値に基づいて順番を設定してもよい。

【数１】

$$Ent_i(D|P) = P(Dp) \times Ent_i(Dp) + P(\overline{Dp}) \times Ent_i(\overline{Dp}) \quad \cdots (1)$$

ただし、$\overline{Dp} = D - Dp$, $P(Dp) = |Dp| / |D|$

【００８５】 ＊ ＊【数２】

$$Ent_i(D) = -(P_i^+ \log P_i^+ + P_i^- \log P_i^-) \quad \cdots (2)$$

ただし、$P_i^+ = |C_i \cap D| / |D|$, $P_i^- = 1 - P_i^+$

【００８６】以上のように、この実施の形態５によれば、所定の順番でパターンを出力するようにしたので、利用者の要求に応じた順番でパターンを出力することができるという効果が得られる。

【００８７】実施の形態６．この発明の実施の形態６による文書情報解析装置は、実施の形態１による文書情報解析装置の制御部３による制御手順を後述のように変更したものである。

【００８８】なお、その他の構成要素は、実施の形態１のものと同様であるので、その説明を省略する。

【００８９】次に動作について説明する。図７は、実施の形態６による文書情報解析装置の動作を説明するフローチャートである。まず、ステップＳＴ１において、文書集合記憶部２は、記録部１に記録された文書情報を読み出し、候補パターン集合生成部６に出力する。

【００９０】ステップＳＴ２において、制御部３は、文書集合記憶部２により文書情報を読み出されたクラスタのうち、後述のステップＳＴ３～ステップＳＴ５、ステップＳＴ８～ステップＳＴ１０、ステップＳＴ６１およびステップＳＴ６３～ステップＳＴ６５の処理を実行されていないクラスタがあるか否かを判定し、すべてのクラスタについてその処理を実行したと判定した場合、パターン抽出の処理を終了し、その処理を実行されていないクラスタが少なくとも１つあると判定した場合、そのクラスタのうちの１つを指定する。

【００９１】候補パターン集合生成部６は、ステップＳＴ３において、供給された文書情報に含まれる各キーワードを１つのキーワードからなる候補パターン（初期候補パターン）とし、クラスタ毎に候補パターン集合（初期候補パターン集合）を生成し、多支持パターン集合生成部４に出力する。そして、ステップＳＴ４において、多支持パターン集合生成部４は、制御部３により指定されたクラスタに属する各候補パターンについて、文書集合記憶部２を制御してそのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップＳＴ５において、そのクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率をそれぞれ計算し、その支持率が所定の最小支持率以上であるパターンを多支持パターンとして選択し、それらの多支持パターンの集合を多支持パターン集合として候補パターン生成部６に出力する。

【００９２】そして、ステップＳＴ６１において、多支持パターン集合生成部４は、多支持パターンを累積多支持パターン集合に追加する。なお、累積多支持パターン集合は、各クラスタについての処理を開始するとき（ステップＳＴ３）において、空集合に設定される。

【００９３】そして、ステップＳＴ８において、制御部３は、多支持パターン集合生成部４により選択されたパターンの数が所定の最小パターン数より小さいか否かを判定し、選択されたパターンの数が所定の最小パターン数より小さくないと判定した場合、ステップＳＴ９に進み、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したか否かを判定する。

【００９４】そして、ステップＳＴ８において選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、または、ステップＳＴ９において選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、ステップＳＴ６３に進み、制御部３は、多支持パターン集合生成部４に制御信号を供給し、そのときに選択されているパターンを高確信度パターン出力部５に出力させる。

【００９５】一方、ステップＳＴ９において、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達していないと判定した場合、制御部３は特に何もしない。したがって、ステップＳＴ１０に進み、候補パターン集合生成部６は、多支持パターン集合生成部４より供給された多支持パターン集合について、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターンの集合を生成し、そのパターン集合を候補パターン集合として多支持パターン集合生成部４に出力する。そして、ステップＳＴ４に戻り、ステップＳＴ１０で生成された候補パターン集合について上述の処理を同様に実行する。

【００９６】このようにして、ステップＳＴ８またはステップＳＴ９の条件が成立するまで、ステップＳＴ４、

ステップＳＴ５、ステップＳＴ６１、ステップＳＴ８～ステップＳＴ１０の処理が反復的に実行される。

【００９７】そして、反復処理が終了した後、ステップＳＴ６３においては、制御部３は、多支持パターン集合生成部４に制御信号を供給し、そのときに選択されているパターンを高確信度パターン出力部５に出力させる。

【００９８】高確信度パターン出力部５は、ステップＳＴ６４において、多支持パターン集合生成部４より供給された各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、多支持パターン集合生成部４より供給されたパターン集合から、確信度が所定の最小確信度以上であるパターンを選択し、ステップＳＴ６５において、そのパターンを図示せぬ装置に出力する。そして、ステップＳＴ２に戻り、制御部３は、次のクラスタに対して同様の処理を行うように各部を制御する。

【００９９】以上のように、この実施の形態６によれば、指定されたクラスタ以外のクラスタにおいてパターンが出現する文書情報を１回だけカウントすればよいので、処理時間を短縮することができるという効果が得られる。

【０１００】なお、この実施の形態６においても、実施の形態５のように、所定の順番でパターンを出力するようにしてもよい。

【０１０１】実施の形態７．この発明の実施の形態７による文書情報解析装置は、実施の形態６による文書情報解析装置の文書集合記憶部２と多支持パターン集合生成部４に後述の機能を追加したものである。

【０１０２】すなわち、実施の形態７における文書集合記憶部２は、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。また、多支持パターン集合生成部４は、候補パターン集合生成部６より供給される候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【０１０３】なお、その他の構成要素は、実施の形態６のものと同様であるので、その説明を省略する。

【０１０４】次に動作について説明する。図８は、実施の形態７による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ５、ステップＳＴ８～ステップＳＴ１０、ステップＳＴ６１、およびステップＳＴ６３～ステップＳＴ６５の処理は、実施の形態６のものと同様であるので、その説明を省略する。

【０１０５】実施の形態７においては、ステップＳＴ１の後のステップＳＴ７１において、文書集合記憶部２が、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。

【０１０６】また、ステップＳＴ６３の前のステップＳＴ７２において、高確信度パターン出力部５が、多支持パターン集合生成部４より供給されるパターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成する。そして、ステップＳＴ６３の処理では、その文書情報が元の文書情報の代わりに使用される。

【０１０７】以上のように、この実施の形態７によれば、パターンを構成する各キーワードについて支持率を計算し、その支持率が所定の最小支持率より低いキーワードをすべての文書情報から除去するようにしたので、パターンが出現するか否かを検査する対象になる文書情報の数を低減させることができ、処理時間を短縮することができるという効果が得られる。

【０１０８】

【発明の効果】以上のように、この発明によれば、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択し、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力し、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するように構成したので、そのクラスタ特有のパターンを抽出することができ

る効果がある。

【０１０９】この発明によれば、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の文書ＩＤを記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の文書ＩＤを読み出し、その文書情報においてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするように構成したので、パターンの出現の検査の対象になる文書情報の数が低減され、処理時間を短縮することができる効果がある。

【０１１０】この発明によれば、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするように構成したので、パターンの出現の検査の対象になるクラスタの数が低減され、処理時間を短縮することができる効果がある。

【０１１１】この発明によれば、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを、すべての文書情報から除去するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【０１１２】この発明によれば、各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【０１１３】この発明によれば、各クラスタについて、初期のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、処理の最初だけにおいて不要なキーワードが除去され、除去の処理にかかる時間を低減することができる効果がある。

【０１１４】この発明によれば、新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去するように構成したので、利用者の必要頻度が低く、かつそのクラスタにしか存在しないような複雑なパターンを除去することができる効果がある。

【０１１５】この発明によれば、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加するように構成したので、キーワード数が多くかつ確信度の低いパターンを除去することができ、キーワード数が少なく確信度の高いパターンが選択され、検索時の効率を向上させることができる効果がある。

【０１１６】この発明によれば、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、その確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力するように構成したので、キーワード数が長くかつ確信度の低いパターンを選択するときの処理を簡略化することができ、処理時間を短縮することができる効果がある。

【０１１７】この発明によれば、クラスタ毎に一括して、所定の順番でパターンを出力するように構成したので、利用者の要求に応じた順番でパターンを出力することができる効果がある。

【０１１８】この発明によれば、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加し、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、多支持

パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するように構成したので、指定されたクラスタ以外のクラスタにおいてパターンが出現する文書情報を１回だけカウントすればよいので、処理時間を短縮することができる効果がある。

【０１１９】この発明によれば、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【０１２０】この発明によれば、すべてのクラスタから選択されたクラスタだけについてパターンを選択し、出力するように構成したので、不要なクラスタに対して処理が行われないので、処理時間を短縮することができる＊効果がある。

【図面の簡単な説明】

【図１】　この発明の実施の形態１による文書情報解析装置の構成を示すブロック図である。

【図２】　図１の文書情報解析装置の動作を説明するフローチャートである。

【図３】　実施の形態２による文書情報解析装置の動作を説明するフローチャートである。

【図４】　実施の形態３による文書情報解析装置の動作を説明するフローチャートである。

【図５】　実施の形態４による文書情報解析装置の動作を説明するフローチャートである。

【図６】　実施の形態５による文書情報解析装置の動作を説明するフローチャートである。

【図７】　実施の形態６による文書情報解析装置の動作を説明するフローチャートである。

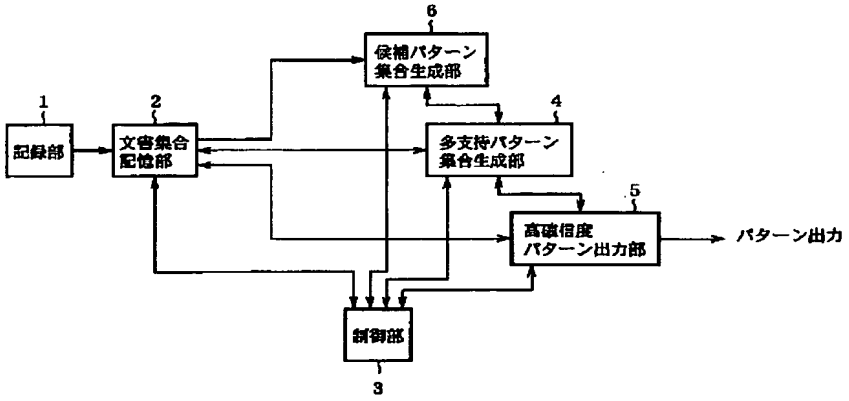
【図８】　実施の形態７による文書情報解析装置の動作を説明するフローチャートである。

【図９】　スポーツ関係の新聞記事を記録したデータベースの一例を示している。

【符号の説明】

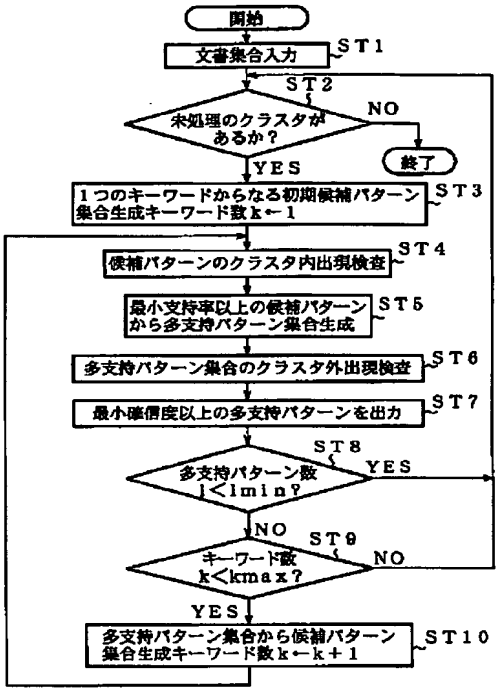
１　記録部、４　多支持パターン集合生成部（第１の選択手段、選択手段）、５　高確信度パターン出力部（第２の選択手段、出力手段）、６　候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）。

【図１】

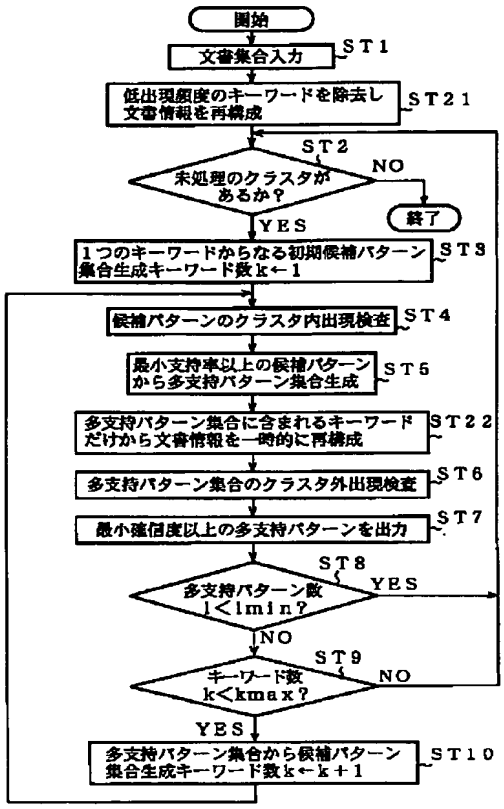


4：多支持パターン集合生成部（第１の選択手段、選択手段）
5：高確信度パターン出力部（第２の選択手段、出力手段）
6：候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）

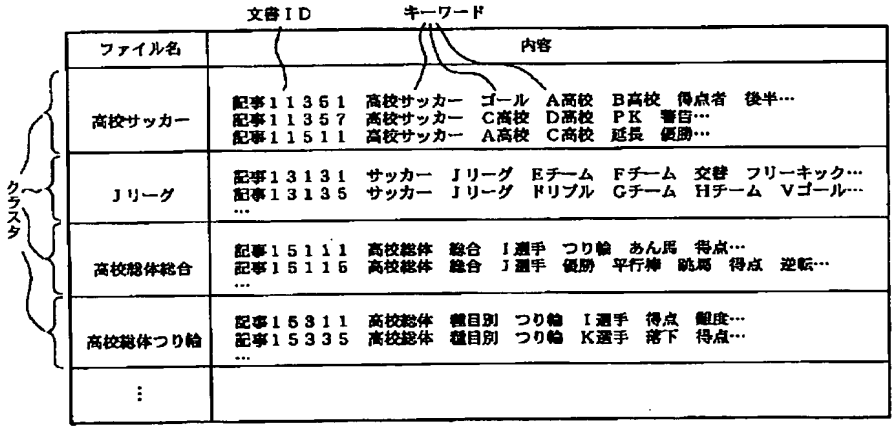
【図２】


開始
文書集合入力
ST1
ST2
未処理のクラスタが
あるか？
NO
終了
YES
１つのキーワードからなる初期候補パターン
集合生成キーワード数k←１
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターン
から多支持パターン集合生成
ST5
多支持パターン集合のクラスタ外出現検査
ST6
最小確信度以上の多支持パターンを出力
ST7
ST8
多支持パターン数
l＜lmin？
YES
NO
ST9
キーワード数
k＜kmax？
NO
YES
多支持パターン集合から候補パターン
集合生成キーワード数k←k＋１
ST10

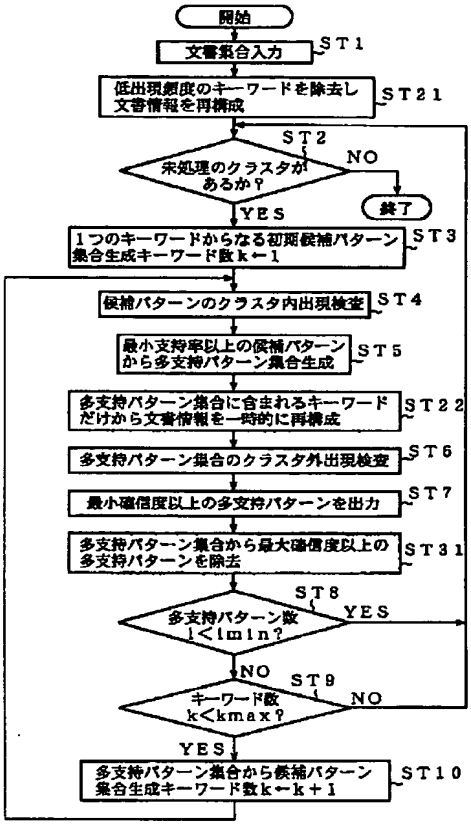
【図３】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し
文書情報を再構成
ST21
ST2
未処理のクラスタが
あるか？
NO
終了
YES
１つのキーワードからなる初期候補パターン
集合生成キーワード数k←１
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターン
から多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワード
だけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
最小確信度以上の多支持パターンを出力
ST7
ST8
多支持パターン数
l＜lmin？
YES
NO
ST9
キーワード数
k＜kmax？
NO
YES
多支持パターン集合から候補パターン
集合生成キーワード数k←k＋１
ST10

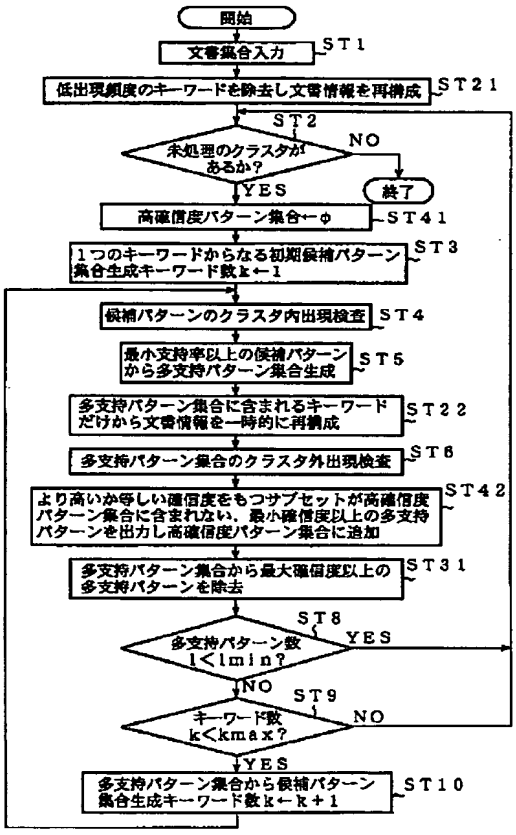
【図９】


文書ＩＤ
キーワード
ファイル名
内容
クラスタ
高校サッカー
記事１１３６１　高校サッカー　ゴール　A高校　B高校　得点者　後半…
記事１１３５７　高校サッカー　C高校　D高校　PK　警告…
記事１１５１１　高校サッカー　A高校　C高校　延長　優勝…
Ｊリーグ
記事１３１３１　サッカー　Ｊリーグ　Eチーム　Fチーム　交替　フリーキック…
記事１３１３５　サッカー　Ｊリーグ　ドリブル　Gチーム　Hチーム　Vゴール…
…
高校総体総合
記事１５１１１　高校総体　総合　Ｉ選手　つり輪　あん馬　得点…
記事１５１１５　高校総体　総合　Ｊ選手　優勝　平行棒　跳馬　得点　逆転…
…
高校総体つり輪
記事１５３１１　高校総体　種目別　つり輪　Ｉ選手　得点　鯉度…
記事１５３３５　高校総体　種目別　つり輪　K選手　落下　得点…
…
⋮

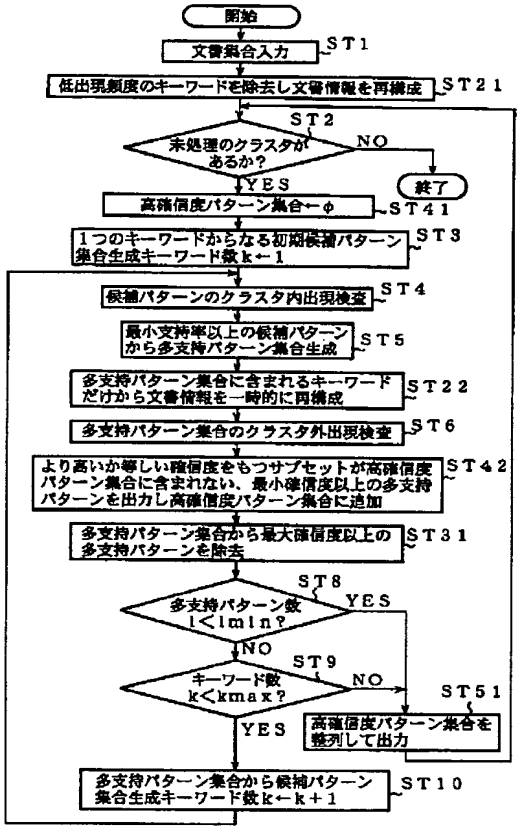
【図４】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
最小確信度以上の多支持パターンを出力
ST7
多支持パターン集合から最大確信度以上の多支持パターンを除去
ST31
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

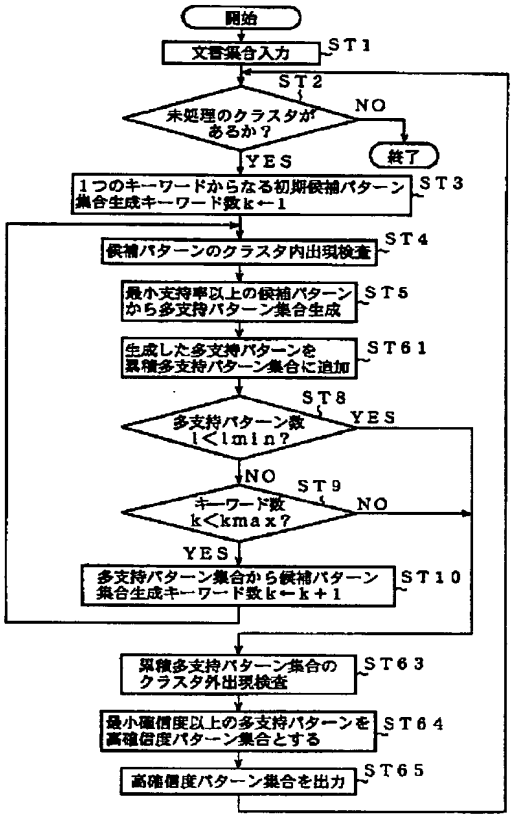
【図５】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
高確信度パターン集合←φ
ST41
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
より高いか等しい確信度をもつサブセットが高確信度パターン集合に含まれない，最小確信度以上の多支持パターンを出力し高確信度パターン集合に追加
ST42
多支持パターン集合から最大確信度以上の多支持パターンを除去
ST31
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

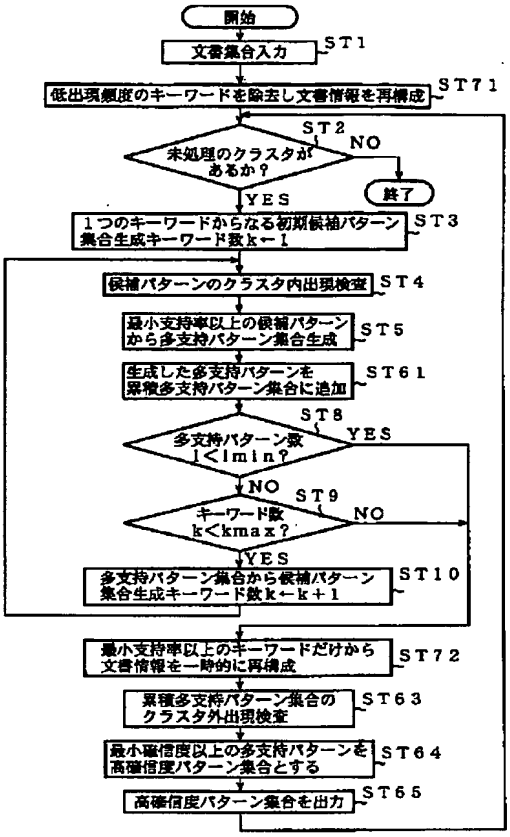
【図６】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
高確信度パターン集合←φ
ST41
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
より高いか等しい確信度をもつサブセットが高確信度パターン集合に含まれない、最小確信度以上の多支持パターンを出力し高確信度パターン集合に追加
ST42
多支持パターン集合から最大確信度以上の多支持パターンを除去
ST31
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
ST51
高確信度パターン集合を整列して出力
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

【図７】


開始
文書集合入力
ST1
ST2
未処理のクラスタがあるか？
NO
終了
YES
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
生成した多支持パターンを累積多支持パターン集合に追加
ST61
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10
累積多支持パターン集合のクラスタ外出現検査
ST63
最小確信度以上の多支持パターンを高確信度パターン集合とする
ST64
高確信度パターン集合を出力
ST65

【図８】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST71
ST2
未処理のクラスタがあるか？
NO
終了
YES
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
生成した多支持パターンを累積多支持パターン集合に追加
ST61
ST8
多支持パターン数l<lmin?
YES
NO
ST9
キーワード数k<kmax?
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10
最小支持率以上のキーワードだけから文書情報を一時的に再構成
ST72
累積多支持パターン集合のクラスタ外出現検査
ST63
最小確信度以上の多支持パターンを高確信度パターン集合とする
ST64
高確信度パターン集合を出力
ST65

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－73415

(43)公開日　平成11年(1999)３月16日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３５０Ｃ

審査請求　未請求　請求項の数４　ＯＬ　（全 11 頁）

(21)出願番号　特願平9－231363

(22)出願日　平成９年(1997)８月27日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町３丁目３番地の１

(72)発明者　田野崎　康雄
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内

(74)代理人　弁理士　須山　佐一

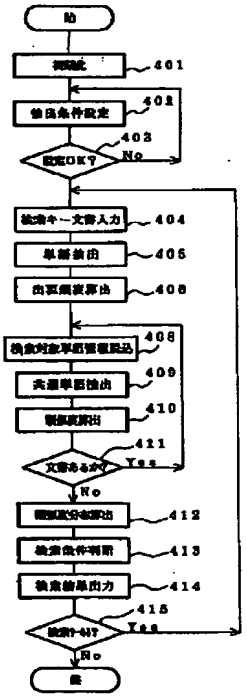
最終頁に続く

(54)【発明の名称】　類似文書検索装置及び類似文書検索方法

(57)【要約】

【課題】　ある文書（検索キー文書）と類似する文書を複数の検索対象文書のなかから検索する装置において、より信憑性の高い類似文書検索を実現する。

【解決手段】　検索キー文書と各検索対象文書との各々の類似度値の統計分布（例えば類似度の平均値）を求め、この統計分布を基準に、ユーザが設定した条件を満足するものを類似文書として抽出する。従来のように単に類似度値が高いものを類似文書として抽出する方式に比べ、類似文書としてより信憑性の高いものを検索結果として得ることができる。また、検索キー文書と各検索対象文書との類似度がどれも一般的な評価基準において高いとは言えないような場合に、類似文書がないことを検索結果として出力する。



1

【特許請求の範囲】

【請求項１】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、

前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、

前記統計情報を基準とする類似文書の抽出条件を設定する抽出条件設定手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度および前記抽出条件設定手段により設定された抽出条件に基づいて類似文書を検索する検索手段とを具備することを特徴とする類似文書検索装置。

【請求項２】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、

前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、

前記統計情報を基準とする類似文書の有無の判定条件を設定する判定条件設定手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度および前記判定条件設定手段により設定された判定条件に基づいて類似文書の有無を判定する判定手段とを具備することを特徴とする類似文書検索装置。

【請求項３】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、

前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、

前記算出された各検索対象文書の類似度の統計情報を求める工程と、

前記統計情報を基準とする類似文書の抽出条件を設定する工程と、

前記算出された各検索対象文書の類似度および前記設定された抽出条件に基づいて類似文書を検索する工程とを具備することを特徴とする類似文書検索方法。

【請求項４】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、

前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、

前記算出された各検索対象文書の類似度の統計情報を求める工程と、

前記統計情報を基準とする類似文書の有無の判定条件を設定する工程と、

前記算出された各検索対象文書の類似度および前記設定された判定条件に基づいて類似文書の有無を判定する工程とを具備することを特徴とする類似文書検索方法。

2

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、電子化された文書データの検索装置に係り、特にある文書データを検索キーとしてこれと類似した文書データを自動検索する類似文書検索装置および類似文書検索方法に関する。

【０００２】

【従来の技術】近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、文書データベース中から指定された文書（以下、検索キー文書と呼ぶ。）に類似する文書の自動検索を行うシステムが実用されてきている。従来の類似文書検索システムでは、検索キー文書に含まれている単語と他の文書（以下、検索対象文書と呼ぶ。）に含まれている単語とを比較し、共通する単語の種類や出現回数・場所などからベクトル空間法により類似度を算出し、最も類似度の高い検索対象文書を検索結果として出力したり、類似度の高い文書から順に出力していた。

【０００３】

【発明が解決しようとする課題】従来の類似文書検索方式は、検索キー文書と文書データベース中の各検索対象文書との類似度を各々算出し、より類似度の高い文書を判定する、例えば最大類似度のものを検索結果として出力している。しかし、その検索結果は必ずしも妥当なものとは言えない。すなわち、上記従来の類似文書検索方式により得た類似度に基づく検索結果は、各検索対象文書に対して求めた各類似度の単純な大小比較により得た検索結果にすぎないため、例えば、検索キー文書と特に類似している検索対象文書が１つも存在しないような場合でも、その中で最も類似度の高い検索対象文書を類似文書として無条件に出力してしまう。

【０００４】本発明はこのような課題を解決するためのもので、複数の検索対象文書のなかから検索キー文書との類似が際立っているものを確実に検索することのできる類似文書検索装置および類似文書検索方法の提供を目的としている。

【０００５】また、本発明は、多くの一般的な類似評価の基準において類似していると呼べる類似文書のみを検索結果として得ることで、信頼性の向上を図ることのできる類似文書検索装置および類似文書検索方法の提供を目的としている。

【０００６】

【課題を解決するための手段】上記目的を達成するために、本発明の類似文書検索装置は、請求項１に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める

統計情報算出手段と、前記統計情報を基準とする類似文書の抽出条件を設定する抽出条件設定手段と、前記類似度算出手段によって算出された各検索対象文書の類似度および前記抽出条件設定手段により設定された抽出条件に基づいて類似文書を検索する検索手段とを具備することを特徴とする。

【０００７】また、本発明の類似文書検索装置は、請求項２に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、前記統計情報を基準とする類似文書の有無の判定条件を設定する判定条件設定手段と、前記類似度算出手段によって算出された各検索対象文書の類似度および前記判定条件設定手段により設定された判定条件に基づいて類似文書の有無を判定する判定手段とを具備することを特徴とする。

【０００８】さらに、本発明の類似文書検索方法は、請求項３に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、前記算出された各検索対象文書の類似度の統計情報を求める工程と、前記統計情報を基準とする類似文書の抽出条件を設定する工程と、前記算出された各検索対象文書の類似度および前記設定された抽出条件に基づいて類似文書を検索する工程とを具備することを特徴とする。

【０００９】さらに、本発明の類似文書検索方法は、請求項４に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、前記算出された各検索対象文書の類似度の統計情報を求める工程と、前記統計情報を基準とする類似文書の有無の判定条件を設定する工程と、前記算出された各検索対象文書の類似度および前記設定された判定条件に基づいて類似文書の有無を判定する工程とを具備することを特徴とする。

【００１０】請求項１および請求項３の発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の抽出条件に基づいて類似文書を検索することで、類似文書としてより妥当性の高いもの、つまり類似度がその他の多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として検索することができる。

【００１１】請求項２および請求項４の発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の有無の判定条件に基づいて類似文書の有無を判定することで、検索キー文書と各検索対象文書との類似度がいずれも多くの一般的な評価基準において高いと言えないような場合に類似文書が存在しないとし、一般的な評価基準において類似していると言える類似文書だけを検索結果として得ることができる。

【００１２】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して詳細に説明する。

【００１３】図１は本発明に係る一実施形態の類似文書検索装置のハードウェア構成を示す図である。

【００１４】同図に示すように、この類似文書検索装置は、ＣＰＵおよびメモリなどから構成される制御装置１、キーボードなどの入力装置２、類似文書の検索結果などを表示する表示装置３、および文書データや類似文書検索のための各文書の単語情報などを格納する外部記憶装置４から構成される。

【００１５】図２に本類似文書検索装置における制御装置１の構成を示す。制御装置１は制御部２００とメモリ部２２９からなる。

【００１６】制御部２００は、初期化部２０１、入力部２０２、出力部２０３、検索対象文書読み出し部２０４、検索対象単語抽出部２０５、検索対象単語出現頻度算出部２０６、検索対象単語情報書込部２０７、検索キー文書入力部２０８、検索キー単語抽出部２０９、検索キー単語出現頻度算出部２１０、検索対象単語情報読み出し部２１１、共通単語抽出部２１２、類似度算出部２１３、類似度統計分布計算部２１４、抽出条件設定部２１５、検索結果出力部２１６などから構成される。　メモリ部２２９は、検索対象文書格納バッファ部２３０、検索対象単語情報格納バッファ部２３１、検索キー文書格納バッファ部２３２、検索キー単語情報格納バッファ部２３４、共通単語情報格納バッファ部２３５、類似度格納バッファ部２３６、抽出条件設定バッファ部２３７、類似度統計分布結果バッファ部２３８、検索結果出力バッファ部２３９などから構成される。

【００１７】初期化部２０１は、上記各バッファ部の初期化を行う。入力部２０２は、ユーザによる入力装置２からの検索キー文書や抽出条件の設定など各種設定の入力を行う。出力部２０３は、入力部２０２により入力された検索キー文書などの各種設定内容を表示装置３に出力する。

【００１８】検索対象文書読み出し部２０４は、外部記憶装置４に格納されている検索対象文書に関する情報を文書データベース化するために、文書データベース化すべき文書情報を外部記憶装置４から読み込み、検索対象文書格納バッファ部２３０に格納する。

【００１９】検索対象単語抽出部２０５は、検索対象文書格納バッファ部２３０に格納されている検索対象文書からの単語の切り出しを行う。そして、切り出した単語のなかからその文書内容を表す上でキーとなる単語を抽出し、抽出した単語種を検索対象単語情報格納バッファ

部２３１に格納する。単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【００２０】検索対象単語出現頻度算出部２０６は、検索対象単語抽出部２０５により抽出された個々のキー単語について、検索対象文書中での出現頻度を算出し、これを検索対象文書の単語情報として検索対象単語情報格納バッファ部２３１に格納する。

【００２１】検索対象単語情報書込部２０７は、検索対象単語情報格納バッファ部２３１に格納されている検索対象文書の単語情報を外部記憶装置４に格納する。

【００２２】検索キー文書入力部２０８は、入力装置２から入力された検索キー文書の情報を検索キー文書格納バッファ部２３２に格納する。

【００２３】検索キー単語抽出部２０９は、検索キー文書格納バッファ部２３２に格納されている検索キー文書からの単語切り出しを行う。そして、その文書の内容を表す上でキーとなる単語を抽出し、抽出した単語種を検索キー単語情報格納バッファ部２３４に格納する。単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【００２４】検索キー単語出現頻度算出部２１０は、検索キー単語抽出部２０９により抽出された個々のキー単語について、検索キー文書中での出現頻度を算出し、これを検索キー文書の単語情報として検索キー単語情報格納バッファ部２３４に格納する。

【００２５】検索対象単語情報読み出し部２１１は、外部記憶装置４に格納されいてる各検索対象文書の単語情報（単語の出現頻度情報）を１文書分ごとに呼び出し、検索対象単語情報格納バッファ部２３１に格納する。

【００２６】共通単語抽出部２１２は、検索キー単語情報格納バッファ部２３４に格納されている検索キー文書の単語情報と検索対象単語情報格納バッファ部２３１に格納されている検索対象文書の単語情報とを比較して、一致する単語の種類と出現頻度情報を共通単語情報格納バッファ部２３５に格納する。

【００２７】類似度算出部２１３は、共通単語情報格納バッファ部２３５に格納されている情報に基づき検索キー文書と検索対象文書との類似度を算出し、その類似度値を類似度格納バッファ部２３６に格納する。

【００２８】類似度統計分布計算部２１４は、類似度格納バッファ部２３６に格納されている検索キー文書と全検索対象文書との類似度値から類似度の平均値や標準偏差値などの統計分布情報を求めて類似度統計分布結果バッファ部２３８に格納する。抽出条件設定部２１５は、入力装置２を介してユーザより入力された、類似度統計結果から類似文書の検索結果を抽出する場合の条件、または類似度統計結果から検索キー文書との類似文書があるとするための条件、または検索キー文書との類似文書がないとするための条件などの抽出条件値を抽出条件設定バッファ部２３７に格納（設定）する。

【００２９】検索結果出力部２１６は、類似度統計分布結果バッファ部２３８に格納されている統計分布情報、抽出条件設定バッファ部２３７に格納されている抽出条件値、さらには類似度格納バッファ部２３６に格納されている各検索対象文書の類似度値から、検索キー文書に対する類似文書検索結果として、検索対象文書の有無、検索対象文書が有る場合の該当文書を判断し、その検索結果を検索結果出力バッファ部２３９に格納し、そして検索結果出力バッファ部２３９の内容を表示装置３に出力する。

【００３０】次に、本実施形態の類似文書検索装置の動作を説明する。

【００３１】最初に、検索対象文書のデータベースの作成手順を図３、図５、図６により説明する。図３はその手順を示すフローチャートである。

【００３２】まず、初期化部２０１により全バッファ部の初期化を行う（ステップ３０１）。続いて検索対象文書読み出し部２０４が、外部記憶措置４から複数のテキスト文書を読み出し、検索対象文書格納バッファ部２３０に検索対象文書として格納する（ステップ３０２）。具体例として、例えば図５に示すような内容のテキスト文書を検索対象文書の一つとして格納したとする。

【００３３】次に、検索対象単語抽出部２０５が、検索対象文書格納バッファ部２３０に格納されている個々の検索対象文書について、形態素解析などによって単語の切り出しを行い、切り出した単語のなかから文書内容を表すキー単語を抽出し、そのキー単語の単語種（例えば品詞情報）を検索対象単語情報格納バッファ部２３１に格納する（ステップ３０３）。

【００３４】次に、検索対象単語出現頻度算出部２０６が、検索対象単語情報格納バッファ部２３１に格納されている検索対象文書のキー単語について、検索対象文書全体での出現頻度を算出し、その結果を検索対象単語情報格納バッファ部２３１に格納する（ステップ３０４）。図６に検索対象単語情報格納バッファ部２３１の格納例を示す。このバッファ部２３１において単語と頻度は対応付けて記述される。例えばキー単語「文書」が文書全体のなかで２回出現している場合は頻度として「２」が記述される。

【００３５】このようにして検索対象単語情報格納バッファ部２３１に格納された情報は、検索対象文書のデータベースとして外部記憶装置４に蓄積される（ステップ３０５）。

【００３６】この後、検索対象文書格納バッファ部２３０に文書データベース化前の検索対象文書が残っているかどうかを判断し（ステップ３０６）、他に検索対象文書があればステップ３０２に戻って、その新たな検索対

象文書についての前記同様の文書データベースの作成が行われる。他に検索対象文書がなければ本処理を終了する。

【００３７】次に、類似文書の検索手順を図４、図７乃至図１６により説明する。図４は類似文書検索手順を示すフローチャートである。

【００３８】まず、初期化部２０１により全バッファ部の初期化を行う（ステップ４０１）。続いて抽出条件設定部２１５が起動される。抽出条件設定部２１５は、入力装置２を通じてユーザより、

１．類似度統計結果から類似文書の検索結果を抽出する場合の条件

２．類似度統計結果から類似文書が存在しているとするための条件

３．類似度統計結果から類似文書が存在していないとするための条件

などの抽出条件値の入力を受け付けて抽出条件設定バッファ部２３７に格納（設定）する（ステップ４０２）。より具体的には、図７に抽出条件設定バッファ部２３７の格納例を示しているように、１．の条件として、「平均類似度の２倍以上の類似度を持つ検索対象文書を検索結果とする。」などが設定される。

【００３９】２．の条件として、「平均類似度の２倍以上の類似度を持つ検索対象文書がある場合」などが設定される。

【００４０】３．の条件として、「すべての検索対象文書の類似度が０．１以下である場合」などが設定される。

【００４１】これらの抽出条件値はユーザにより任意に決定される。

【００４２】この抽出条件の設定が完了したら（ステップ４０３）、検索キー文書入力部２０８が起動される。検索キー文書入力部２０８は、入力装置２を通じてユーザより検索キー文書の入力を受け付け、入力した検索キー文書の情報を検索キー文書格納バッファ部２３２に格納する（ステップ４０４）。具体例として、図８に示すような検索キー文書が入力されてバッファ部２３２に格納されたとする。

【００４３】次に、検索キー単語抽出部２０９が、検索キー文書格納バッファ部２３２に格納されている検索キー文書から形態素解析などによって単語の切り出しを行い、切り出した単語のなかから文書内容を表すキー単語を抽出し、そのキー単語の単語種（例えば品詞情報）を検索キー単語情報格納バッファ部２３４に格納する（ステップ４０５）。

【００４４】次に、検索キー単語出現頻度算出部２１０が、検索キー単語情報格納バッファ部２３４に格納されている個々のキー単語について、検索キー文書全体のなかでの出現頻度を算出し、これを検索キー単語情報格納バッファ部２３４にキー単語と対応付けて格納する（ステップ４０６）。図９に検索キー単語情報格納バッファ部２３４の格納例を示す。このバッファ部２３４においてキー単語と頻度は対応付けて記述され、例えばキー単語「今後」が２回出現している場合は頻度として「２」が記述される。

【００４５】次に、検索対象単語情報読み出し部２１１が、外部記憶装置４に格納されている各検索対象文書の単語情報を１文書ごとに読み込み、検索対象単語情報格納バッファ部２３１に書き込む（ステップ４０８）。

【００４６】この後、共通単語抽出部２１２が起動され、共通単語抽出部２１２は、検索対象単語情報格納バッファ部２３１と検索キー単語情報格納バッファ部２３４とに共通に格納されているキー単語を検出し、共通単語情報格納バッファ部２３５に格納する（ステップ４０９）。例えば、図１０に示すように、図６の検索対象単語情報格納バッファ部２３１と図９の検索キー単語情報格納バッファ部２３４に共通するキー単語として「画像」が検出され、このキー単語「画像」とその頻度情報「３」を共通単語情報格納バッファ部２３５に対応付けて格納する。

【００４７】次に、類似度算出部２１３が、共通単語情報格納バッファ部２３５に格納されている情報に基づき検索キー文書と検索対象文書との類似度をベクトル空間法などにより算出し、その類似度値を類似度格納バッファ部２３６に格納する（ステップ４１０）。例えば、図１１に示すような各検索対象文書ごとの類似度が類似度格納バッファ部２３６に格納される。

【００４８】すべての検索対象文書について類似度計算が完了すると（ステップ４１１）、類似度統計分布計算部２１４が起動する。類似度統計分布計算部２１４は、類似度格納バッファ部２３６に格納されている各検索対象文書の類似度の統計分布を算出し、その結果を類似度統計分布結果バッファ部２３８に格納する（ステップ４１２）。例えば、図１２に示すように、各検索対象文書の類似度の平均値「０．２５」を類似度の統計分布情報として求まる。

【００４９】次に、検索結果出力部２１６が、抽出条件設定バッファ部２３７に格納されている抽出条件値を判断し（ステップ４１３）、その抽出条件値、類似度統計分布結果バッファ部２３８に格納されている統計分布情報、さらには類似度格納バッファ部２３６に格納されている各検索対象文書の類似度値から、検索キー文書に対する類似文書の検索結果として、検索対象文書の有無や、検索対象文書が有る場合の該当文書を判断し、その結果を検索結果出力バッファ部２３９に格納する。　例えば、１．の条件（類似度統計結果から類似文書の検索結果を抽出する場合の条件）に対し、図１３に示すように、上記条件を満足する検索対象文書があれば、そのＩＤを検索結果出力バッファ部２３９に格納する。

【００５０】また、２．の条件（類似度統計結果から検

索キー文書との類似文書が存在しているとするための条件）に対し、該条件を満足している場合は、図１４に示すように、当該検索キー文書を類似文書有りの検索キー文書として、そのＩＤを検索結果出力バッファ部２３９に格納する。

【００５１】さらに、３．の条件（類似度統計結果から検索キー文書との類似文書が存在していないとするための条件）に対し、該条件を満足する場合は、図１５に示すように、当該検索キー文書を類似文書無しの検索キー文書として、そのＩＤを検索結果出力バッファ部２３９に格納する。

【００５２】検索結果出力部２１６は、検索結果出力バッファ部２３９の内容を、例えば図１６に示すような形式で表示装置３に出力する（ステップ４１４）。図１６の例では、例えば図１４に示すＩＤ「２」の類似文書有りの検索キー文書と類似する検索対象文書としてＩＤ「１」「４２」「５４」「３１４」などがあることを示している。また、２．と３．の各条件により、類似する検索対象文書がないと判断された検索キー文書に対しては類似文書がないことが表示される。

【００５３】この後、次の検索キー文書がある場合はステップ４０４に戻ってその検索キー文書を入力し、以降同様の処理を行う。次の検索キー文書がなければ本処理を終了する。

【００５４】なお、本動作例では、１つの検索キー文書に対する検索結果を得たところでこれを表示するようにしたが、すべての検索キー文書に対する検索結果を得た後、検索結果を見たい検索キー文書を指定すると、その検索結果が表示されるようにしてもよい。

【００５５】このように本実施形態の類似文書検索装置においては、検索キー文書と各検索対象文書との各々の類似度値の統計分布（この実施形態では類似度の平均値）を求め、この統計分布を基準に、ユーザが設定した条件を満足するものを類似文書として抽出することで、従来のように単に類似度値が高いものを類似文書として抽出する方式に比べ、類似文書としてより信憑性の高いものを検索結果として得ることができる。すなわち、本実施形態は、類似度の統計分布を基準としているので、検索キー文書との類似度がその他多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として得られる。

【００５６】また、従来の方式では、検索キー文書と各検索対象文書との類似度が、どれも一般的な評価基準において高いとは言えないような場合でも検索結果として類似文書を無条件に出力してしまうが、本実施形態では、このような場合において類似文書がないことを検索結果として出力する。このような点からも、本実施形態の類似文書検索装置によれば、類似文書として信憑性の高い検索結果を得ることができ、類似文書検索効率を大幅に高めることができる。

【００５７】

【発明の効果】以上説明したように本発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の抽出条件に基づいて類似文書を検索することで、類似文書としてより妥当性の高いもの、つまり類似度がその他の多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として得ることができる。

【００５８】また、本発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の有無の判定条件に基づいて類似文書の有無を判定することで、検索キー文書と各検索対象文書との類似度がいずれも一般的な評価基準において高いと言えないような場合に類似文書が存在しないとし、一般的な評価基準において類似していると言える類似文書だけを検索結果として得ることができる。

【図面の簡単な説明】

【図１】本発明に係る一実施形態の類似文書検索装置のハードウェア構成を示す図

【図２】図１の類似文書検索装置における制御装置の機能ブロック図

【図３】検索対象文書のデータベースの作成手順を示す図

【図４】類似文書検索手順を示す図

【図５】検索対象文書の例

【図６】検索対象単語情報の格納例

【図７】抽出条件の設定例

【図８】検索キー文書の例

【図９】検索キー単語情報の例

【図１０】共通単語情報の例

【図１１】各検索対象文書の類似度の例

【図１２】類似度の平均値の例

【図１３】類似文書の検索結果の例

【図１４】類似文書の検索結果の例

【図１５】類似文書の検索結果の例

【図１６】類似文書の検索結果の出力例

【符号の説明】

２００……制御部
２１０……初期化部
２０２……入力部
２０３……出力部
２０４……検索対象文書読み出し部
２０５……検索対象単語抽出部
２０６……検索対象単語出現頻度算出部
２０７……検索対象単語情報書込部
２０８……検索キー文書入力部
２０９……検索キー単語抽出部
２１０……検索キー単語出現頻度算出部
２１１……検索対象単語情報読み出し部

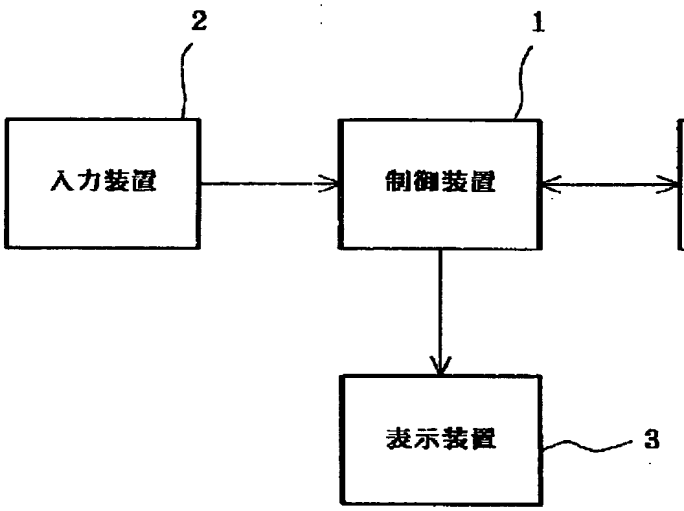
11

２１２……共通単語抽出部
２１３……類似度算出部
２１４……類似度統計分布計算部
２１５……抽出条件設定部
２１６……検索結果出力部
２２９……メモリ部
２３０……検索対象文書格納バッファ部
２３１……検索対象単語情報格納バッファ部

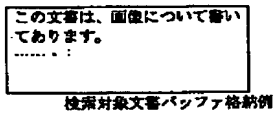
12

＊２３２……検索キー文書格納バッファ部
２３３……検索キー単語情報格納バッファ部
２３５……共通単語情報格納バッファ部
２３６……類似度格納バッファ部
２３７……抽出条件設定バッファ部
２３８……類似度統計分布結果バッファ部
２３９……検索結果出力バッファ部

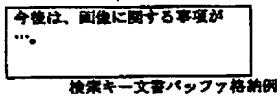
【図１】



【図５】

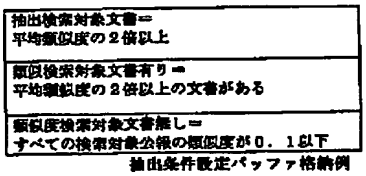


【図８】

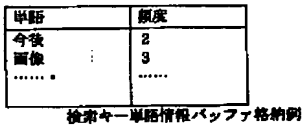


【図６】

【図７】

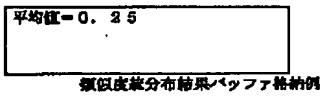


【図９】

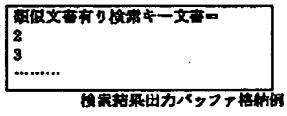


【図１０】

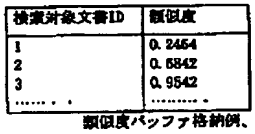
【図１２】



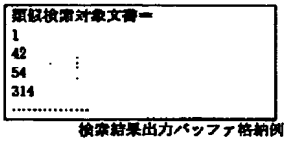
【図１４】



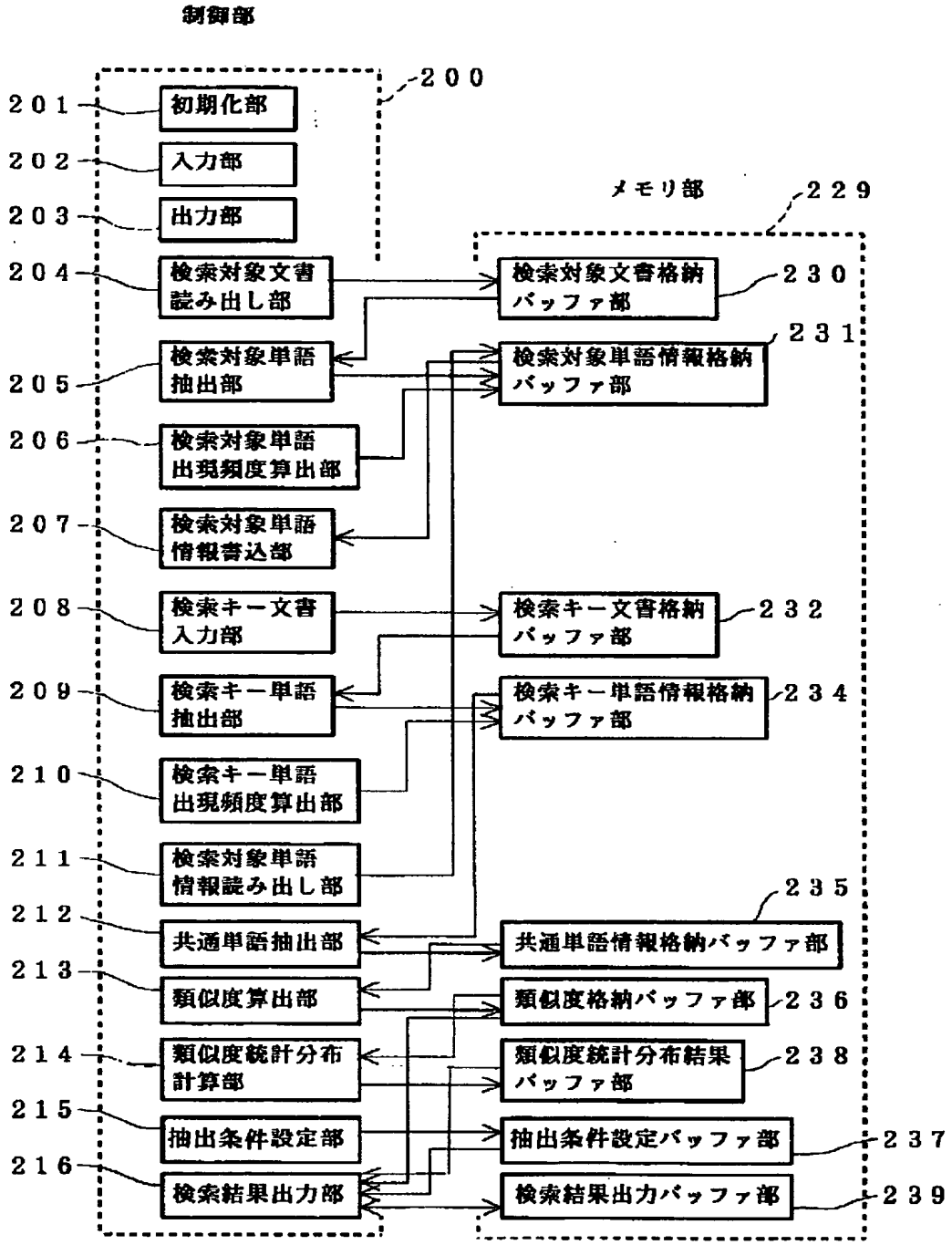
【図１１】



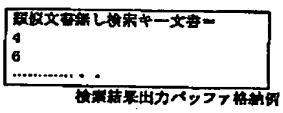
【図１３】



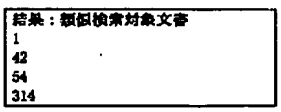
【図２】


制御部
２００
２０１
初期化部
２０２
入力部
２０３
出力部
２０４
検索対象文書
読み出し部
２０５
検索対象単語
抽出部
２０６
検索対象単語
出現頻度算出部
２０７
検索対象単語
情報書込部
２０８
検索キー文書
入力部
２０９
検索キー単語
抽出部
２１０
検索キー単語
出現頻度算出部
２１１
検索対象単語
情報読み出し部
２１２
共通単語抽出部
２１３
類似度算出部
２１４
類似度統計分布
計算部
２１５
抽出条件設定部
２１６
検索結果出力部
メモリ部
２２９
検索対象文書格納
バッファ部
２３０
検索対象単語情報格納
バッファ部
２３１
検索キー文書格納
バッファ部
２３２
検索キー単語情報格納
バッファ部
２３４
２３５
共通単語情報格納バッファ部
類似度格納バッファ部
２３６
類似度統計分布結果
バッファ部
２３８
抽出条件設定バッファ部
２３７
検索結果出力バッファ部
２３９

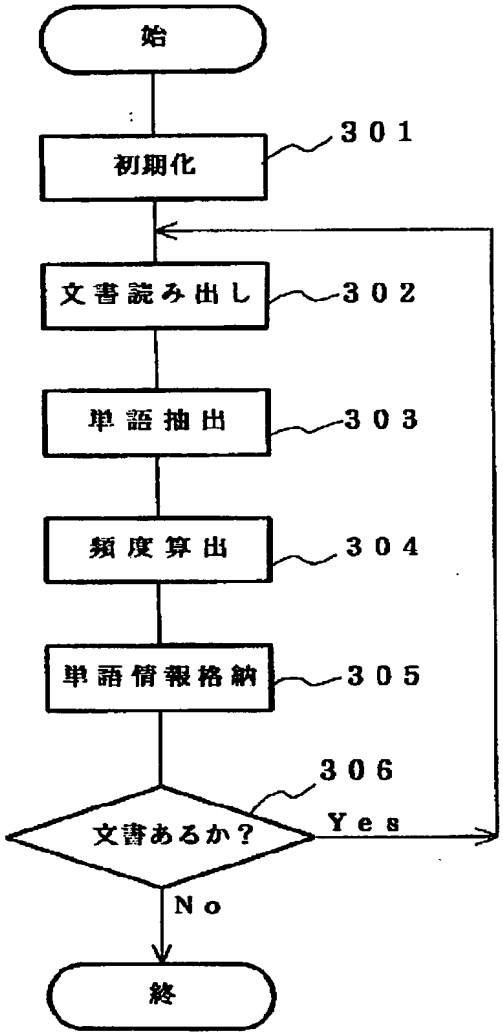
【図１５】


類似文書探し検索キー文書ー
4
6
検索結果出力バッファ格納例

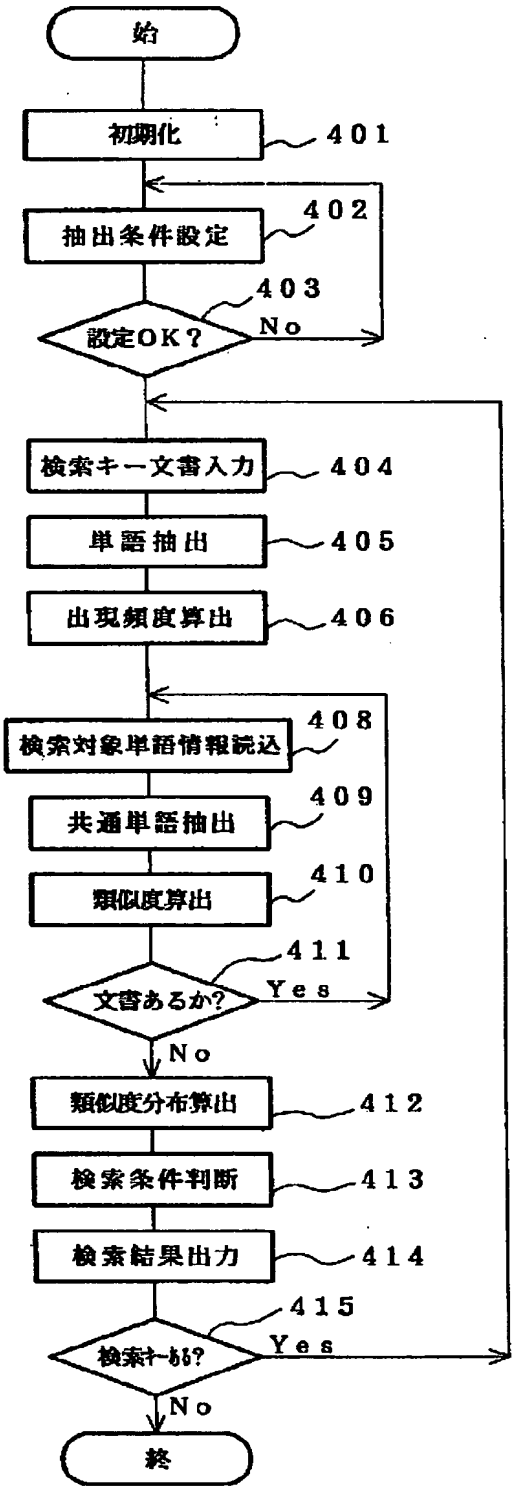
【図１６】


結果：類似検索対象文書
1
42
54
314

【図３】


始
初期化
３０１
文書読み出し
３０２
単語抽出
３０３
頻度算出
３０４
単語情報格納
３０５
３０６
文書あるか？
Ｙｅｓ
Ｎｏ
終

【図４】


始
初期化
401
抽出条件設定
402
設定OK？
403
No
検索キー文書入力
404
単語抽出
405
出現頻度算出
406
検索対象単語情報読込
408
共通単語抽出
409
類似度算出
410
文書あるか?
411
Yes
No
類似度分布算出
412
検索条件判断
413
検索結果出力
414
検索キーあり?
415
Yes
No
終

フロントページの続き

(72)発明者　中本　幸夫
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　仁科　卓哉
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　久保田　直秀
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－345239

(43)公開日　平成11年(1999)12月14日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ａ
			15/40	３７０Ａ

審査請求　未請求　請求項の数11　ＯＬ　（全 20 頁）

(21)出願番号　特願平10－153231

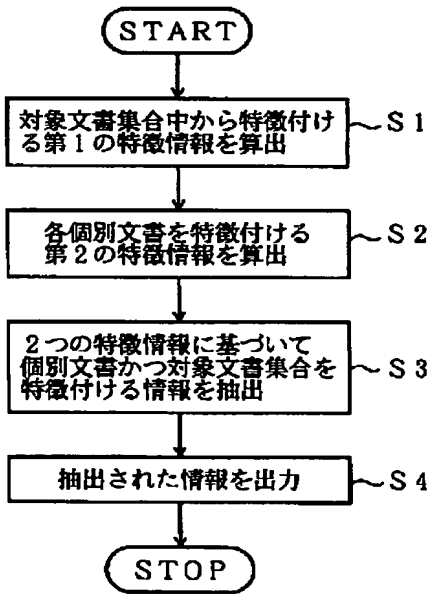
(22)出願日　平成10年(1998)６月２日

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　巌寺　俊哲
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　弁理士　伊東　忠彦

(54)【発明の名称】　文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体

(57)【要約】
【課題】　予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供する。
【解決手段】　本発明は、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出し、対象文書集合中の各個別文書を他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出し、第１の特徴情報と、前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、抽出された前記情報を各個別文書を特徴づける情報として出力する。

本発明の原理を説明するための図



1

【特許請求の範囲】

【請求項１】　文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、

前記対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出し、

前記対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出し、

前記第１の特徴情報と、前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、

抽出された前記情報を各個別文書を特徴づける情報として出力することを特徴とする文書情報抽出方法。

【請求項２】　文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、

前記対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出手段と、

前記対象文書集合中の各個別文書について、他の個別文書を特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出手段と、

前記第１の特徴情報算出手段で算出された前記第１の特徴情報と、前記第２の特徴情報算出手段で算出された前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出手段と、

抽出された前記情報を各個別文書を特徴づける情報として出力する特徴情報出力手段とを有することを特徴とする文書情報抽出装置。

【請求項３】　前記入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、

前記標準文書集合更新手段に与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析手段と、

前記標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、

前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、

前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析手段と、

前記各対象文書集合中の単語と該単語の出現頻度を各文書を対応付けて記憶する対象文書集合解析結果記憶手段

2

と、

前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、

前記対象文書集合全体としての特徴情報を、前記対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、

前記対象文書集合全体特徴算出手段によって算出された、前記対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、

前記対象文書集合中の各個別文書の特徴情報を、前記個別文書解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、

前記個別文書特徴算出手段によって算出された、前記対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、

前記個別文書解析結果記憶手段または、前記対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、

前記対象文書集合解析結果記憶手段または、前記標準文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する基準情報一時記憶手段と、

前記目的情報一時記憶手段に記憶されているデータと前記基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、

前記対象文書集合全体特徴記憶手段に記憶されているデータと前記個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報算出手段と、

前記特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、

前記特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、

前記特徴表現抽出手段により前記各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、

前記特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する請求項２記載の文書情報抽出装置。

【請求項４】　前記特徴情報算出手段は、

前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項３記載の文書情報抽出装置。

【請求項５】　前記特徴情報算出手段は、

前記特徴情報を算出する際にχ２乗検定を用いる請求項４記載の文書情報抽出装置。

【請求項６】　前記特徴表現抽出手段は、

前記対象文書集合中の各文書中の全単語に、単語の特徴

を数値化した特徴情報スコアを付与する特徴スコア付与手段と、
前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出手段と、
前記平均算出手段により求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定手段とを含む請求項３記載の文書情報抽出装置。
【請求項７】　文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、
前記対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出プロセスと、
前記対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出プロセスと、
前記第１の特徴情報算出プロセスで算出された前記第１の特徴情報と、前記第２の特徴情報算出プロセスで算出された前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出プロセスと、
抽出された前記情報を各個別文書を特徴づける情報として出力する特徴情報出力プロセスとを有することを特徴とする文書情報抽出プログラムを格納した記憶媒体。
【請求項８】　前記入力記憶手段から標準文書集合を受け取る標準文書集合更新プロセスと、
前記標準文書集合更新プロセスに与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析プロセスと、
前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、
前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析プロセスと、
前記対象文書集合全体としての特徴情報を、前記各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、
前記対象文書集合中の各個別文書の特徴情報を、前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書集合解析結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、
前記個別文書の解析結果または、前記対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、前記対象文書集合解析プロセスまたは、前記標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、
前記対象文書集合全体特徴算出プロセスによって算出された、前記対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、前記個別文書特徴算出プロセスによって算出された、前記対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、
前記特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、
前記特徴表現抽出プロセスにより前記各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する請求項７記載の文書情報抽出プログラムを格納した記憶媒体。
【請求項９】　前記特徴情報算出プロセスは、
前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項８記載の文書情報抽出プログラムを格納した記憶媒体。
【請求項１０】　前記特徴情報算出プロセスは、
前記特徴情報を算出する際にχ２乗検定を用いる請求項９記載の文書情報抽出プログラムを格納した記憶媒体。
【請求項１１】　前記特徴表現抽出プロセスは、
前記対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、
前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出プロセスと、
前記平均算出プロセスにより求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定プロセスとを含む請求項８記載の文書情報抽出プログラムを格納した記憶媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、文書情報抽出方法

及び装置及び文書情報抽出プログラムを格納した記憶媒体に係り、特に、文書情報処理に用いられる文書情報抽出処理において、特定の話題についての文書集合中の各文書より、当該話題に関連し、かつ、文書集合中の他文書との相違をより明確に示す情報を抽出する文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体に関する。

【０００２】

【従来の技術】近年、インターネットが急速に普及している。さらに、データ記憶装置は、大容量化、低価格化している。これに伴って大量で多様な情報が、ネットワークを介して容易に利用可能になっている。また、ＷＷＷの普及と共に多くのユーザが相互に情報を生成し、利用している。しかし、情報洪水と言われるように利用できる情報量が飛躍的に増加するに従って、これらの情報の中から有益な情報を見つけ出して取捨選択することが困難になってきている。

【０００３】このような大量の情報を全て閲覧し、有益な情報を探索し、選別することは困難である。従って、適切な情報を効率的に利用するためには、大量の情報から特徴的な情報を抽出し、必要十分な情報を選択的に利用可能にする必要がある。現在、情報を選択的に利用する手段として情報検索技術が用いられている。しかし、ネットワークを介して利用できる情報は、その分量が膨大であり、その内容も多岐に渡っている。このため、一度の検索結果として多くの類似の情報を含む文書が選択されてしまう。

【０００４】さらに、検索結果から適切な情報を含む文書を選択することを支援するために、多くの場合、各文書を特徴付ける部分情報を文書から抽出し、各検索結果文書に付与し、利用者に提示している。膨大な数の文書の各文書から適量・適質な特徴的な部分情報の、人手による抽出は、困難である。また、人手により作業を行った場合、抽出される情報は、作業者のもっている主観や知識に影響されるため、複数の作業者によって抽出が行われると抽出された情報の品質を均質に保つことができない。そこで、情報を選択的に利用するためには、文書内容から適切な部分情報を自動的に決定し、抽出する情報抽出技術が必要となる。

【０００５】従来の情報抽出技術の代表的なものとして次の２つが挙げられる。第１には、パターンマッチングによる情報抽出であり、第２には、ヒューリスティックスを用いた情報抽出技術がある。パターンマッチングによる情報抽出技術は、ある単語列をパターンとして予め保持しておき、パターンマッチング処理によって情報を抽出する技術である。これは、特定の情報が、限られたパターンによって表現されることが多いという考え方に基づいている。

【０００６】例えば、予め、「＜メーカ＞は、＜製品＞の販売を開始した」のようなパターンを用意しておくことにより、文書中からこのパターンにマッチする「○×コンピュータは新型コンピュータの販売を開始した。」という文を抽出する。また、ヒューリスティックスを用いた情報抽出技術は、文の文書中での位置情報、タイトルや見出しに出現する単語、手がかり語句の有無を組み合わせて文の重要度を判定し、重要と判定された文を抽出する技術である。

【０００７】

【発明が解決しようとする課題】しかしながら、上記従来のパターンマッチングによる情報抽出技術は、必要とするパターンを予め用意しておくことが必要である。このため、パターンマッチングしない新たな情報の抽出を行うことができないという問題がある。また、同一の文書からは、必ず同一の部分が抽出される。

【０００８】また、上記従来のヒューリスティックスを用いた情報抽出技術は、前述のパターンマッチングによる情報抽出技術と同様に、予めヒューリスティックスを用意しておくことが必要である。このため、新たな情報の抽出には、不向きである。また、同一の文書からは、必ず同一の部分が抽出される。さらに、ある文書タイプで有効なヒューリスティックスが別の文書タイプで有効であるとは限らない。例えば、新聞記事などでは、位置情報が有効である。また、学術論文では、手がかり語句によるヒューリスティックスが有効である。インターネット上には、様々なタイプの文書が混在しており、文書タイプを自動的に判定することが必要となる。しかし、現状では、文書タイプを判別する有効な技術がない、という問題がある。

【０００９】また、タイトルや見出しが存在しない文書や、手がかり語句が殆ど出現しない文書も多いため、ヒューリスティックスが有効に働かないことが多いという問題がある。本発明は、上記の点に鑑みなされたもので、予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供することを目的とする。

【００１０】

【課題を解決するための手段】図１は、本発明の原理を説明するための図である。本発明（請求項１）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出し（ステップ１）、対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出し（ステップ２）、第１の特徴情報と、第２の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書か

ら抽出し（ステップ３）、抽出された情報を各個別文書を特徴づける情報として出力する（ステップ４）。

【００１１】図２は、本発明の原理構成図である。本発明（請求項２）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出手段１と、対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出手段２と、第１の特徴情報算出手段１で算出された第１の特徴情報と、第２の特徴情報算出手段２で算出された第２の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出手段３と、抽出された情報を各個別文書を特徴づける情報として出力する特徴情報出力手段４とを有する。

【００１２】本発明（請求項３）は、入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、標準文書集合更新手段に与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析手段と、標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析手段と、各対象文書集合中の単語と該単語の出現頻度を対応付けて記憶する対象文書集合解析結果記憶手段と、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、対象文書集合全体としての特徴情報を、対象文書集合解析結果記憶手段及び標準文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、対象文書集合全体特徴算出手段によって算出された、対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、対象文書集合中の各個別文書の特徴情報を、個別文書解析結果記憶手段及び対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、個別文書特徴算出手段によって算出された、対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、個別文書解析結果記憶手段または、対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、対象文書集合解析結果記憶手段または、標準文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する基準情報一時記憶手段と、目的情報一時記憶手段に記憶されているデータと基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、対象文書集合全体特徴記憶手段に記憶されているデータと個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報算出手段と、特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、特徴表現抽出手段により各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する。

【００１３】本発明（請求項４）は、特徴情報算出手段において、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項５）は、特徴情報算出手段において、特徴情報を算出する際にχ２乗検定を用いる。本発明（請求項６）は、特徴表現抽出手段において、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与手段と、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出手段と、平均算出手段により求められた平均の値が最大の文書内部分表現を文書の特徴表現として抽出する特徴表現決定手段とを含む。

【００１４】本発明（請求項７）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出プロセスと、対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出プロセスと、第１の特徴情報算出プロセスで算出された第１の特徴情報と、第２の特徴情報算出プロセスで算出された第２の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出プロセスと、抽出された情報を各個別文書を特徴づける情報として出力する特徴情報出力プロセスとを有する。

【００１５】本発明（請求項８）は、入力記憶手段から標準文書集合を受け取る標準文書集合更新プロセスと、

標準文書集合更新プロセスに与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析プロセスと、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の出現頻度を算出する対象文書集合解析プロセスと、対象文書集合全体としての特徴情報を、各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、対象文書集合中の各個別文書の特徴情報を、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書解析結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、個別文書の解析結果または、対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、対象文書集合解析プロセスまたは、標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、対象文書集合全体特徴算出プロセスによって算出された、対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、個別文書特徴算出プロセスによって算出された、対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、特徴表現抽出プロセスにより各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する。

【００１６】本発明（請求項９）は、特徴情報算出プロセスにおいて、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項１０）は、特徴情報算出プロセスにおいて、特徴情報を算出する際にχ２乗検定を用いる。本発明（請求項１１）は、特徴表現抽出プロセスにおいて、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出プロセスと、平均算出プロセスにより求められた平均の値が最大の文書内部分表現を文書の特徴表現として抽出する特徴表現決定プロセスとを含む。

【００１７】上述のように、本発明は、対象文書集合を標準文書集合に対して特徴付ける情報を対象文書集合中から算出し、次に、対象文書集合中の各個別文書について、他の個別文書に対してその文書を特徴付ける情報を個別文書から算出し、これらの処理により得られた特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を各個別文書から抽出する。抽出された情報を各個別文書を特徴付ける情報として出力する。

【００１８】これにより、例えば、「桜の花見」について検索を実行した結果、得られる文書集合を「対象文書集合」とし、この文書検索装置が検索対象とする文書集合全体を「標準文書集合」とした場合を想定すると、この場合、対象文書集合を標準文書集合に対して特徴付ける情報は、「桜の花見」に関する情報、例えば、桜の見頃や名所に関する情報になる。

【００１９】また、各個別文書を他の個別文書に対して特徴付ける情報は、他の文書と差異がある情報、例えば、「４月の上旬」のような時間的な情報や、「上野公園」のような場所的情報になる。これらの情報を組み合わせて用いることにより、各個別文書からその文書を特徴付ける表現、例えば、「…見頃は、４月上旬…」、「…名所：上野公園…」のような時間的、場所的記述を表している部分を抽出することが可能となる。

【００２０】このように、本発明では、大量の文書情報から各文書を特徴付ける情報を適時に適質適量抽出可能となる。また、予め情報抽出に使用する知識等を用意する必要がなく、様々な文書から情報抽出に適用可能である。

【００２１】

【発明の実施の形態】図３は、本発明の文書情報抽出装置の構成を示す。同図に示す構成は、例えば、文書検索システムの一部を構成し、検索された文書集合中の各文書から各文書間の相違を明確に示し、各文書を特徴付ける適切な情報を抽出し、提示することにより、使用者が必要とする文書を選択することを支援する装置である。以下、文書検索システム本体から出力された文書集合中の各文書から情報抽出する場合を想定して説明する。ここで、当該文書情報抽出装置は、入力として文書集合を受信し、予め、提供されている初期情報を用いて入力された文書集合を処理し、各文書毎に情報を抽出し、出力するものである。

【００２２】同図に示す構成は、監視制御部１０、入力記憶装置２０、転送媒体３０、各種制御処理を実行する処理部１０１～１１０及び各種データを記憶する記憶部２０１～２０９から構成される。処理部は、標準文書集合更新部１０１、標準文書集合解析部１０２、対象文書

集合入力部１０３、対象文書集合解析部１０４、対象文書集合全体特徴算出部１０５、個別文書特徴算出部１０６、特徴情報算出部１０７、特徴表現抽出部１０８、特徴表現出力部１０９、目的情報特徴スコア算出部１１０から構成される。

【００２３】記憶部は、標準文書集合解析結果記憶部２０１、対象文書集合解析結果記憶部２０２、対象文書集合全体特徴記憶部２０３、個別文書解析結果記憶部２０４、個別文書特徴記憶部２０５、特徴情報記憶部２０６、特徴表現記憶部２０７、基準情報一時記憶部２０８、及び目的情報一時記憶部２０９から構成される。各処理部１０１～１１０を総合的に監視制御する監視制御部１０に、処理部１０１～１０９と、記憶部２０１～２０７が接続される。また、対象文書集合全体特徴算出部１０５と個別文書特徴算出部１０６には、目的情報特徴スコア算出部１１０、基準情報一時記憶部２０８及び目的情報一時記憶部２０９が接続される。さらに、基準情報一時記憶部２０８、目的情報一時記憶部２０９は、目的特徴スコア算出部１１０にも接続される。

【００２４】ここで、各処理部は、例えば、ディジタル電子計算機で構成され、それぞれＣＰＵと、動作プログラムとそれを実行するためのデータを記録するＲＯＭと、ワーキングメモリとして用いられるＲＡＭとを備える。なお、全処理部を１つのディジタル電子計算機で構成してもよい。さらに、各記憶部２０１～２０９は、例えば、ハードディスクメモリなどのメモリに記憶される。

【００２５】また、入力記憶部２０には、本装置に与えられる標準文書集合、対象文書集合が一定の順序で記憶されている。入力記憶部２０は、半導体メモリ装置、あるいは、ハードディスクやフロッピーディスクによって実現することができる。転送媒体３０には、本装置の処理結果が与えられる通信チャネルまたは記録媒体である。

【００２６】以下の説明において、「標準文書集合」とは、特徴情報抽出の対象となり得る文書全体集合、または、文書全体集合を母集合とし、その標本集合となる文書集合を指す。また、「対象文書集合」とは、特徴情報抽出の対象の文書集合を指す。さらに、「個別文書」とは、「対象文書集合」中の各文書を指す。

【００２７】「特徴表現」とは、各「個別文書」から抽出される表現であり、「対象文書集合」中の他の文書に対して、当該「個別文書」を特徴付ける表現である。標準文書集合更新部１０１は、入力記憶装置２０から標準文書集合を受け取る。標準文書集合解析部１０２は、標準文書集合更新部１０１に与えられた標準文書集合中の各文書を解析し、その文を構成する単語とその単語の標準文書集合中での出現頻度を算出する。

【００２８】標準文書集合解析結果記憶部２０１は、標準文書集合中の単語とその単語の出現頻度を対応付けて

記憶する。対象文書集合入力部１０３は、入力記憶装置２０から複数の文書で構成される対象文書集合を受け取る。対象文書集合解析部１０４は、対象文書集合中の各文書を解析し、その各個別文書を構成する単語と当該単語の出現頻度を算出する。

【００２９】対象文書集合解析結果記憶部２０２は、対象文書集合解析部１０４で求められた各対象文書中の単語と単語の出現頻度を各文書と対応付けて記憶する。個別文書解析結果記憶部２０４は、対象文書集合解析部１０４で求められた対象文書集合中の各個別文書中の単語とその単語の出現頻度をその単語が出現した文書と対応付けて記憶する。

【００３０】対象文書集合全体特徴算出部１０５は、対象文書集合全体としての特徴情報を、上記対象文書集合解析結果記憶部２０２に記憶されている情報及び上記標準文書集合解析結果記憶部２０１に記憶されている情報を用いて算出する。対象文書集合全体特徴記憶部２０３は、対象文書集合全体特徴算出部１０５によって算出された、対象文書集合全体としての特徴情報を記憶する。

【００３１】個別文書特徴算出部１０６は、対象文書集合中の各個別文書の特徴情報を個別文書解析結果記憶部２０４に記憶されている情報及び対象文書集合解析結果記憶部２０２に記憶されている情報を用いて算出する。個別文書特徴記憶部２０５は、個別文書特徴算出部１０６によって算出された対象文書集合中の各個別文書の特徴情報を記憶する。

【００３２】目的情報一時記憶部２０９は、個別文書解析結果または、対象文書集合解析結果を一時的に記憶する。基準情報一時記憶部２０８は、対象文書集合解析結果、または、標準文書集合解析結果を一時的に記憶する。目的情報特徴スコア算出部１１０は、目的情報一時記憶部２０９に記憶されているデータと上記基準情報一時記憶部２０８に記憶されているデータを比較し、上記の目的情報一時記憶部２０９に記憶されているデータの特徴スコアを算出する。

【００３３】特徴情報算出部１０７は、対象文書集合全体特徴記憶部２０３に記憶されているデータと個別文書特徴記憶部２０５に記憶されているデータを用いて各個別文書の特徴情報を算出する。特徴情報記憶部２０６は、特徴情報算出部１０７において算出された各個別文書の特徴情報を記憶する。

【００３４】特徴表現抽出部１０８は、特徴情報記憶部２０６に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する。特徴表現記憶部２０７は、特徴表現抽出部１０８により各個別文書から抽出された特徴表現を記憶する。特徴表現出力部１０９は、特徴表現記憶部２０７に記憶されている特徴表現を転送媒体３０に与える。

【００３５】

【実施例】以下、図面と共に本発明の実施例を説明す

る。まず、監視制御部１０及び対象文書集合全体特徴算出部１０５、個別文書特徴算出部１０６に接続される記憶部２０１～２０９について説明する。標準文書集合解析結果記憶部２０１は、標準文書集合解析部１０２の処理結果である、標準文書集合の解析結果を記憶・保持する。標準文書集合の解析結果とは、標準文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現及び各単語出現頻度を対応付けたものである。解析結果は、単語表現、出現頻度の２つのカラムからなるテーブルとして表現・記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【００３６】対象文書集合解析結果記憶部２０２は、対象文書集合解析部１０４の処理結果の１つとして得られる対象文書集合の解析結果を記憶・保持する。対象文書集合とは、本装置が接続される情報検索装置において検索作業の実行を結果として得られる文書集合である。対象文書集合の解析結果とは、対象文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現と対象文書集合中での各単語の出現頻度を対応付けたものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の２つのカラムからなるテーブルとして、表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【００３７】個別文書解析結果記憶部２０４は、対象文書集合解析部１０４の処理結果の１つとして得られる個別文書の解析結果を記憶・保持する。ここで、個別文書とは、本装置が接続される情報検索装置において、検索作業の実行の結果として得られる文書集合、即ち、前述した対象文書集合に含まれる各文書である。個別文書の解析結果とは、対象文書集合中の各文書毎に記述されている文章を形態素解析し、各単語の表現、及びその文書中での各単語の出現頻度を対応付け、文書毎に記録したものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の２つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。また、対象文書集合中の各文書毎に１個のテーブルが構成される。

【００３８】対象文書集合全体特徴記憶部２０３は、対象文書集合全体特徴算出部１０５の処理結果として得られる、標準文書集合に対する対象文書集合全体として特徴を点数化した情報を記憶・保持する。ここで、対象文書集合全体特徴とは、標準文書集合中の単語の出現頻度分布と対象文書集合中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、標準文書集合中の出現頻度分布と対象文書集合中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。対象文書集合中で特徴的な単語、即ち、標準文書集合中の出現頻度分布と対象文書集合中の出現分布の相違が大きい単語ほど大きな数値をとる。対象文書集合全体特徴は、各単語表現とその単語の出現分布の特徴を数値化し、表現した特徴スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。対象文書集合全体に対して１つのテーブルが対応する。

【００３９】個別文書特徴記憶部２０５は、個別文書特徴算出部１０６の処理結果として得られる。対象文書集合全体に対する対象文書集合中の各個別文書の特徴を点数化し、記憶・保持する。ここで、個別文書特徴とは、対象文書集合中の単語の出現頻度分布と対象文書集合中の各個別文書中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、対象文書集合中の出現分布と個別文書中の出現分布の相違が大きい単語ほど、大きな数値をとる。即ち、個別文書（対象文書集合中の各文書）に特徴的な単語ほどより大きな数値を持つ。個別文書特徴は、各単語表現とその単語の出現分布の特徴を数値化し表現した特徴スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。また、個別文書毎に１つのテーブルが対応する。

【００４０】特徴情報記憶部２０６は、特徴情報算出部１０７の処理結果として得られる、各文書の特徴情報をその情報の算出元である文書と対応付けて記憶・保持する。ここで、特徴情報とは、各単語毎に対象文書集合中での特徴スコアと個別文書中での特徴スコアをかけた数値であり、対象文書集合中に特徴的単語であり、かつ個別文書中でも特徴的な単語ほど大きな数値を持つ。特徴情報は、各単語表現とその単語の特徴を数値化した特徴スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴スコアの対応を表す。また、個別文書毎に１つのテーブルが対応する。

【００４１】特徴表現記憶部２０７は、特徴表現抽出部１０８の処理結果として得られる、対象文書集合中の各文書毎の特徴表現をその情報の抽出元である文書と対応付けて記憶・保持する。各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた部分構造を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列を構成する単語全体の特徴スコアの平均が最大の部分である。

【００４２】基準情報一時記憶部２０８は、目的情報特徴スコア算出部１１０における特徴スコアの算出に用いる基準情報を一時的に記憶・保持する。基準情報は、単語表現、出現頻度の２つのカラムからなるテーブルとし

て表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部２０８には、一度に上記の形式のテーブルが１個、記憶・保持される。

【００４３】目的情報一時記憶部２０９は、目的情報特徴スコア算出部１１０における特徴スコアの算出において、特徴スコアの算出の対象となる目的情報を一時的に記憶・保持する。目的情報は、単語表現、出現頻度の２つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部２０９には、一度に上記形式のテーブルが一個、記憶・保持される。

【００４４】次に、図３に示す各処理部について説明する。監視制御部１０は、処理部１０１～１０９を制御し、データフローを統制するモジュールである。図４は、本発明の一実施例の文書情報抽出処理のフローチャートである。以下、同図のフローチャートに沿って、各処理部の動作を説明する。

【００４５】ステップ１０１）　監視制御部１０において、標準文書集合が更新されているか否かが判断される。更新された場合には、ステップ１０２に移行し、更新されていない場合は、ステップ１０５に移行する。

ステップ１０２）　監視制御部１０は、更新された標準文書集合を入力記憶装置２０から標準文書集合更新部１０１へ転送する。

【００４６】ステップ１０３）　この時点で、標準文書集合更新部１０１は、転送された標準文書集合に対して、標準文書集合更新処理を実行し、処理結果である標準文書集合更新結果を監視制御部１０に出力する。監視制御部１０は、標準文書集合更新部１０１から出力されたすべての標準文書集合更新結果を標準文書集合解析部１０２へ転送する。

【００４７】ステップ１０４）　標準文書集合解析部１０２は、標準文書集合解析処理を実行し、処理結果を監視制御部１０へ出力する。これにより、監視制御部１０は、標準文書集合解析部１０２から出力されたすべての標準文書集合解析結果を標準文書集合解析結果記憶部２０１に転送し、その内容を更新し、新たに転送された値を記憶・保持する。

【００４８】ステップ１０５）　監視制御部１０によって対象文書集合が入力されたか否かが判断される。入力された場合は、ステップ１０６へ移行し、入力されていない場合には、ステップ１０５の処理を繰り返す。

ステップ１０６）　監視制御部１０は、入力された対象文書集合を入力記憶装置２０から対象文書集合入力部１０３へ転送する。対象文書集合入力部１０３は、入力された対象文書集合に対して対象文書集合入力処理を実行し、処理結果を監視制御部１０に出力する。監視制御部１０は、対象文書集合入力部１０３から出力される対象文書集合入力処理結果を対象文書集合解析部１０４へ転送する。

【００４９】ステップ１０７）　対象文書集合解析部１０４は、対象文書集合解析処理を実行し、解析結果を監視制御部１０へ出力する。

ステップ１０８）　監視制御部１０は、対象文書集合解析部１０４から出力された対象文書集合解析結果を対象文書集合解析結果記憶部２０２に転送すると共に、個別文書解析結果を個別文書解析結果記憶部２０３に転送し、各記憶部の内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、標準文書集合解析結果記憶部２０１に記憶されている標準文書集合解析結果を対象文書集合全体特徴算出部１０５へ転送すると共に、対象文書集合解析結果記憶部２０２に記憶されている対象文書集合解析結果を対象文書集合全体特徴算出部１０５へ転送する。

【００５０】ステップ１０９）　対象文書集合全体特徴算出部１０５は、監視制御部１０から転送された標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴算出処理を実行し、処理結果を監視制御部１０に出力する。監視制御部１０は、対象文書集合全体特徴算出部１０５から出力された対象文書集合全体特徴を対象文書集合全体特徴記憶部２０３に転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、対象文書集合解析結果記憶部２０２に記憶されている対象文書集合解析結果を個別文書特徴算出部１０６へ転送すると共に、個別文書解析結果記憶部２０４に記憶されている個別文書集合解析結果を個別文書特徴算出部１０６へ転送する。

【００５１】ステップ１１０）　個別文書特徴算出部１０６は、転送されてきた対象文書集合解析結果と個別文書解析結果に基づいて個別文書特徴算出処理を実行する。処理結果である、個別文書特徴情報は、監視制御部１０に出力する。監視制御部１０は、個別文書特徴算出部１０６から出力された個別文書特徴情報を個別文書特徴記憶部２０５に転送し、その内容を更新し、新たに転送された値を記憶・保持する。監視制御部１０は、対象文書集合全体特徴記憶部２０３に記憶されている対象文書集合全体特徴を特徴情報算出部１０７へ転送する。それと共に、個別文書特徴記憶部２０５に記憶されている個別文書特徴を特徴情報算出部１０７へ転送する。

【００５２】ステップ１１１）　特徴情報算出部１０７は、対象文書集合全体特徴と個別文書特徴に基づいて特徴情報算出処理を実行し、処理結果である特徴情報を監視制御部１０へ出力する。監視制御部１０は、特徴情報算出部１０７から出力された特徴情報を特徴情報記憶部２０６へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、特徴

情報記憶部２０６に記憶されている特徴情報を特徴表現抽出部１０８へ転送する。

【００５３】ステップ１１２）　特徴表現抽出部１０８は、転送されてきた特徴情報に基づいて、特徴表現抽出処理を実行し、処理結果である特徴表現を、監視制御部１０へ出力する。監視制御部１０は、特徴表現抽出部１０８から出力された特徴表現を特徴表現記憶部２０７へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、特徴表現記憶部２０７に記憶されている特徴表現を特徴表現出力部１０９へ転送する。

【００５４】ステップ１１３）　特徴表現出力部１０９は、転送されてきた特徴表現について特徴表現出力処理を実行する。処理結果は、監視制御部１０へ出力され、監視制御部１０において、特徴情報出力部１０９から出力された特徴情報出力処理結果を転送媒体３０へ出力する。

ステップ１１４）　すべての処理が終了か否かを判定し、すべての処理が終了している場合には、当該監視制御処理を終了する。また、終了していない場合には、ステップ１０５に移行し、上述の処理を繰り返す。

【００５５】以下に、各部の詳細な処理について説明する。標準文書集合更新部１０１では、監視制御部１０から転送された標準文書集合に対して標準文書集合更新処理が実行される。この処理は、以降の処理の前処理であり、入力された標準文書集合中の各文書から本装置による処理に必要のない部分を除去する。また、以降の処理で対応している文字コードへ変換される。処理結果は、監視制御部１０へ出力される。

【００５６】標準文書集合解析部１０２では、監視制御部１０から転送される標準文書集合更新処理結果に対して標準文書集合解析処理が実行される。この処理は、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び転送されてきた標準文書集合中での各単語の出現頻度を対応付けて記録するものである。解析結果は、単語表現、出現頻度の２つのカラムからなるテーブルとして監視制御部１０へ出力される。このテーブルにおいて、各行には、各単語表現、及びその単語の出現頻度が記述される。また、このテーブルは、各単語表現をキーとして対応する全ての行を検索できる構造をとる。

【００５７】対象文書集合入力部１０３では、監視制御部１０から転送されてくる対象文書集合に対して対象文書集合入力処理が実行される。この処理は、以降の処理の前処理であり、転送された対象文書集合中の各文書から本装置による処理に必要のない部分を除去する。また、以降の処理で対応している文字コードへ変換される。処理結果は監視制御部１０へ出力される。

【００５８】対象文書集合解析部１０４では、監視制御部１０から転送されてくる対象文書集合入力処理結果に対して対象文書集合解析処理が実行され、処理結果として対象文書集合解析結果、個別文書解析結果が監視制御部１０へ出力される。対象文書集合解析処理は、まず、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び、その文書中での各単語出現頻度を対応付け、文書毎のテーブルに記録する。このテーブルは、対象文書集合中のすべての文書に対して、文書毎に１個作られる。これらのテーブルが、個別文書解析結果として監視制御部１０へ出力される。次に、これらのすべてのテーブルに対して同一単語表現の出現頻度が合算され、対象文書集合全体に対して、１個のテーブルが作られる。このテーブルが対象文書集合解析結果として監視制御部１０へ出力される。上記のすべてのテーブルは、単語表現を記録するカラムと、その単語の出現頻度を記録するカラムの２つのカラムから構成される。また、このテーブルは、各単語表現をキーとして対応するすべての行を検索できる構造をとる。

【００５９】対象文書集合全体特徴算出部１０５では、監視制御部１０から転送されてくる、標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴を算出する。対象文書集合全体特徴の算出手順は、次のようになる。

①　標準文書集合解析結果を基準情報一時記憶部２０８に転送する。

【００６０】②　対象文書集合解析結果を目的情報一時記憶部２０９に転送する。

③　目的情報特徴スコア算出部１１０を起動し、算出結果が返されるのを待機する。

④　目的情報特徴スコア算出部１１０から算出結果が返されると、それを対象文書集合全体特徴として監視制御部１０へ出力する。

【００６１】個別文書特徴算出部１０６では、監視制御部１０から転送されてくる、対象文書集合解析結果と個別文書解析結果に基づいて、対象文書集合中に含まれている各文書毎に個別文書特徴を算出する。個別文書特徴の算出手順は、以下のようになる。

①　対象文書集合解析結果を基準情報一時記憶部２０８に転送する。

【００６２】②　個別文書解析結果を目的情報一時記憶部２０９に転送する。

③　目的情報特徴スコア算出部１１０を起動し、算出結果が返されるのを待機する。

④　目的情報特徴スコア算出部１１０から算出結果が返されると、それを個別文書特徴として監視制御部１０へ出力する。

【００６３】特徴情報算出部１０７では、監視制御部１０から転送されてくる、対象文書集合特徴と個別文書特徴に基づいて、対象文書集合中の各文書毎に、特徴情報を算出する。算出結果は、監視制御部１０へ出力される。特徴情報は、各単語表現とその単語の特徴を数値化

した特徴情報スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴情報スコアの対応を表す。

【００６４】また、個別文書毎に１つのテーブルが構成される。特徴情報算出の手順は、次のようになる。

① 各単語毎に対象文書集合特徴中での特徴スコアと個別文書特徴中での特徴スコアを読み出す。

② これらのスコアの積を求める。この結果が各単語の特徴情報スコアとなる。

【００６５】対象文書集合中に特徴的単語であり、かつ、個別文書中でも特徴的な単語ほど、この特徴情報スコアは大きな数値を持つ。特徴表現抽出部１０８では、監視制御部１０から転送されてくる特徴情報を用いて、対象文書集合中の各文書から各文書毎にその文書に特徴的な表現（特徴表現）を抽出し、その表現が抽出された文書と対応付け、監視制御部１０へ出力する。

【００６６】各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造（例えば、段落）を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列に含まれる各単語の特徴情報スコアの平均が最大の部分である。

【００６７】特徴表現として、各文書から一文を抽出する場合、特徴表現抽出の手順は以下のようになる。

① まず、文書中の全単語に特徴情報スコアを付与する。

② 次に、文書中の各文毎に、その文を構成する単語に付与されている特徴情報スコアの平均を求める。

【００６８】③ 特徴情報スコアの平均が最大の文をその文書の特徴表現として抽出する。特徴表現出力部１０９では、監視制御部１０から転送されてくる、各個別文書に対応する特徴表現を監視制御部１０を通して転送媒体３０に出力する。目的情報特徴スコア算出部１１０では、基準情報一時記憶部２０８と目的情報一時記憶部２０９に各々記憶・保存されている基準情報と目的情報に基づいて、特徴スコアを算出し、出力する。

【００６９】特徴スコアとは、基準情報中の単語の出現頻度分布と目的情報中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、基準情報中の出現頻度分布と目的情報中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。即ち、目的情報に特徴的な単語ほど、より大きな数値を持つ。

【００７０】特徴スコアは、各単語毎にその単語の出現頻度分布に対して、χ２乗検定の考え方を用いて算出する。χ２乗検定は、「いくつかの群で、ある変数の分布に差があるかどうか」を検定することができる。本発明では、この変数を文書中の単語とする。例えば、標準文書集合に対する対象文書集合の特徴スコアを算出する場合、対象文書集合と標準文書集合中の全単語の出現総数

と対象文書集合中の全単語の出現総数と標準文書集合中の全単語の出現総数から計算される各単語の各文書集合中での出現頻度の期待値の分布と、実際に観測される各単語の各文書集合中での出現頻度の分布からχ２乗値を算出する。この値が大きくなるほど分布に差があることになり、そのような単語ほど偏って出現していることになる。本発明では、この値を用いて各単語の特徴スコアを算出する。

【００７１】以下に標準文書集合と対象文書集合が与えられている場合の具体的な例を用いて説明する。図５は、本発明の一実施例の標準文書集合の一部の例を示し、図６は、本発明の一実施例の対象文書集合の一部の例を示す。以下に各処理部における詳細な処理動作を前述のフローチャートに基づいて説明する。

【００７２】まず、標準文書集合が更新されているか否かが判定される（ステップ１０１）。この例では更新されていると判明したものとする。図５に示す標準文書集合が、標準文書集合更新部１０１に転送される（ステップ１０２）。なお、標準文書集合は、文書集合であり、その文書数は十分に大きいことが望ましい。標準文書集合更新部１０１では、標準文書集合中の各文書から以後の処理に不要である部分が除去される（ステップ１０３）。例えば、ＨＴＭＬ形式の文書の場合は、ＨＴＭＬタグが除去される。また、ワープロ文書の場合は、文字飾り等が除去される。さらに、文書を構成している文字のコードがまちまちである場合は、１つのコードに統一される。この結果は、監視制御部１０に出力される。監視制御部１０は、これを標準文書集合解析部１０２に転送する。

【００７３】標準文書集合解析部１０２は、転送されてきた標準文書集合を解析する（ステップ１０４）。即ち、各文書毎にその文書が記述している文章を形態素解析し、標準文書集合中の単語表現とその単語の出現頻度を求める。図７は、本発明の一実施例の標準文書集合解析結果の例を示す。同図に示す標準文書集合解析結果は、監視制御部１０に出力される。監視制御部１０は、標準文書集合解析結果を標準文書集合解析結果記憶部２０１に転送する。標準文書集合解析結果記憶部２０１は、転送されてきた標準文書集合解析結果を記憶・保持する。

【００７４】以上により、標準文書集合に関する処理が完了する。次に、対象文書集合が入力されているか否かが判定される（ステップ１０５）。入力されている場合は、以下のように処理が進行する。図６に示す対象文書集合が、対象文書集合入力部１０３に入力される（ステップ１０６）。対象文書集合入力部１０３は、入力された文書集合から以降の処理に不要の部分を除去する。また、以降の処理に対応する文字コードへ変換する。処理結果は、監視制御部１０に出力される。監視制御部１０は、対象文書集合入力部１０３から出力された対象文書

集合を対象文書集合解析部１０４に転送する。

【００７５】対象文書集合解析部１０４は、転送された対象文書集合を解析する（ステップ１０７）。解析結果は対象文書集合解析結果と個別文書解析結果である。対象文書集合解析結果は、対象文書集合を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図８に示す。また、個別文書解析結果は、対象文書集合を構成する各文書毎のその文書を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図９に示す。これらの結果は監視制御部１０に出力される。

【００７６】監視制御部１０は、対象文書集合解析結果を対象文書集合解析結果記憶部２０２に、個別文書解析結果を個別文書解析結果記憶部２０４にそれぞれ転送する。対象文書集合解析結果記憶部２０２は、転送されてきた対象文書集合解析結果を記憶・保持する。同様に、個別文書解析結果記憶部２０４は、転送されてきた個別文書解析結果を記憶・保持する（ステップ１０８）。

【００７７】次に、対象文書集合全体特徴を計算する（ステップ１０９）。まず、監視制御部１０は、標準文書集合解析結果記憶部２０１に記憶・保持されている標準文書集合解析結果と、対象文書集合解析結果記憶部２０２に記憶・保持されている対象文書集合解析結果を対象文書集合全体特徴算出部１０５に転送する。対象文書集合全体特徴算出部１０５は、転送されてきた標準文書集合解析結果を基準情報一時記憶部２０８に、同様に、転送されてきた対象文書集合解析結果を目的情報一時記憶部２０９に転送する。

【００７８】次に、特徴スコア算出部１１０を起動する。特徴スコア算出部１１０では、基準情報一時記憶部２０８と目的情報一時記憶部２０９を参照して特徴スコアを算出する。算出結果は、対象文書集合全体特徴算出部１０５に出力させる。図１０は、特徴スコアの算出結果を示す。対象文書集合全体特徴算出部１０５では、特徴スコア算出部１１０の出力結果を対象文書集合全体特徴として図１０に示すような結果を監視制御部１０に出力する。監視制御部１０は、対象文書集合全体特徴を対象文書集合全体特徴記憶部２０４に転送する。対象文書集合全体特徴記憶部２０４は、転送されてきた対象文書全体特徴を記憶・保持する。

【００７９】次に、対象文書集合中の各文書毎に、個別文書特徴を計算する（ステップ１１０）。まず、監視制御部１０は、対象文書集合解析結果記憶部２０２に記憶・保持されている対象文書集合解析結果を個別文書特徴算出部１０６に転送する。また、監視制御部１０は、個別文書解析結果記憶部２０４に記憶・保持されている個別文書解析結果を、一文書分毎に個別文書特徴算出部１０６に転送する。

【００８０】次に、特徴スコア算出部１１０を起動する。特徴スコア算出部１１０では、基準情報一時記憶部２０８と目的情報一時記憶部２０９を参照して特徴スコアを算出する。算出結果は、個別文書特徴算出部１０６に出力される。その結果の一部を図１１に示す。個別文書特徴算出部１０６では、目的特徴スコア算出部１１０の出力結果を個別文書特徴として監視制御部１０に出力する。その結果一部を図１１に示す。監視制御部１０は、個別文書特徴算出部１０６で得られた個別文書特徴を個別文書特徴記憶部２０５に転送する。個別文書特徴記憶部２０５は、転送されてきた個別文書特徴を記憶・保持する。

【００８１】次に、対象文書集合中の各文書毎に特徴情報を算出する（ステップ１１１）。まず、監視制御部１０は、対象文書集合全体特徴記憶部２０３に記憶・保持されている対象文書集合全体特徴を特徴情報算出部１０７に転送する。また、同様に、個別文書特徴記憶部２０５に記憶・保持されている個別文書特徴を各文書毎に特徴情報算出部１０７に転送する。

【００８２】特徴情報算出部１０７は、監視制御部１０から転送されてきた対象文書全体特徴と個別文書特徴を用いて特徴情報を算出し、算出結果を監視制御部１０へ出力する。算出結果の一部を図１２に示す。ここで、特徴情報とは、各単語毎に対象文書集合中での特徴スコアと個別文書中での特徴スコアを掛けた数値であり、対象文書集合中に特徴的単語であり、かつ、個別文書でも特徴的な単語程大きな数値を持つ。

【００８３】監視制御部１０は、特徴情報算出部１０７の出力結果を各文書毎に特徴情報記憶部２０６に転送する。特徴情報記憶部２０６は、転送されてきた特徴情報を文書毎に記憶・保持する。次に、対象文書集合中の各文書から特徴表現を抽出する（ステップ１１２）。まず、監視制御部１０は、特徴情報記憶部２０６に記憶・保持されている特徴情報を特徴表現抽出部１０８に転送する。

【００８４】特徴表現抽出部１０８では、監視制御部１０から転送されてきた特徴情報を用いて各文書から特徴表現を抽出し、その結果を監視制御部１０へ出力する。抽出結果を図１３に示す。監視制御部１０は、特徴表現抽出部１０８の出力結果を各文書毎に特徴表現記憶部２０７に転送する。

【００８５】特徴表現記憶部２０７は、監視制御部１０から転送されてきた特徴表現を文書毎に記憶・保持すると共に、監視制御部１０に出力する。監視制御部１０は、特徴表現を特徴表現出力部１０９に転送する。監視制御部１０は、対象文書集合中のすべての文書において、特徴表現抽出を終了後、特徴表現出力部１０９は、特徴表現記憶部２０７に記憶・保持されている特徴表現をそれが抽出された文書と対応付けし、転送媒体３０に出力する（ステップ１１３）。

【００８６】以上の実施例において、種々の定義値を用いているが、これらの値は設計値であり、下記のように

23

必要に応じて変更してもよい。
・特徴情報の算出にχ２乗検定の考え方を用いているが、他の手法で算出してもよい。
・特徴量スコアの算出単位として単語を用いたが、この単位は文字や一定長の文字列でもよい。
【００８７】また、上記の実施例では、図３に示す構成に基づいて説明しているが、この例に限定されることなく、専用のハードウェア回路によって実現することも可能であり、さらに、プログラムされたコンピュータによって実現することも可能である。つまり、監視制御部１０、標準文書集合更新部１０１、標準文書集合解析部１０２、対象文書集合入力部１０３、対象文書集合解析部１０４、対象文書集合全体特徴算出部１０５、個別文書特徴算出部１０６、特徴情報算出部１０７、特徴表現抽出部１０８、特徴表現出力部１０９、目的情報特徴スコア算出部１１０をプログラムとして構築し、文書情報抽出装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスク、ＣＤ－ＲＯＭ等の可搬記憶媒体に格納しておき、本発明を実施する際に、インストールすることにより、容易に本発明を実現することが可能である。
【００８８】なお、本発明は、上記の実施例に限定されることなく、特許請求の範囲内で種々変更・応用が可能である。
【００８９】
【発明の効果】上述のように、本発明によれば、予め情報抽出知識やパターンなどを用意することなく、文書情報を抽出することが可能となる。これにより、使用開始時に想定したものと対象とする文書内容に差異が生じた場合や、新たな情報を含んでいる場合においても、適切に文書情報を抽出することが可能である。
【００９０】また、文書集合中の各文書を比較するのに適した各文書を特徴付ける文書情報が抽出できるので、文書検索システムの出力編集装置に適用することにより、効率的に検索結果の文書集合から文書を選択、閲覧することができる。
【図面の簡単な説明】
【図１】本発明の原理を説明するための図である。
【図２】本発明の原理構成図である。
【図３】本発明の文書情報抽出装置の構成図である。
【図４】本発明の一実施例の文書情報抽出処理のフローチャートである。

24

＊【図５】本発明の一実施例の標準文書集合の例である。
【図６】本発明の一実施例の対象文書集合の例である。
【図７】本発明の一実施例の標準文書集合解析結果の例である。
【図８】本発明の一実施例の対象文書集合解析結果の例である。
【図９】本発明の一実施例の個別文書解析結果の例である。
【図１０】本発明の一実施例の対象文書集合全体特徴の例である。
【図１１】本発明の一実施例の個別文書特徴の例である。
【図１２】本発明の一実施例の特徴情報の例である。
【図１３】本発明の一実施例の特徴表現の例である。
【符号の説明】
１　第１の特徴情報算出手段
２　第２の特徴情報算出手段
３　個別文書特徴抽出手段
４　特徴情報出力手段
１０　監視制御部
２０　入力記憶装置
３０　転送媒体
１０１　標準文書集合更新部
１０２　標準文書集合解析部
１０３　対象文書集合入力部
１０４　対象文書集合解析部
１０５　対象文書集合全体特徴算出部
１０６　個別文書特徴算出部
１０７　特徴情報算出部
１０８　特徴表現抽出部
１０９　特徴表現出力部
１１０　目的情報スコア算出部
２０１　標準文書集合解析結果記憶部
２０２　対象文書集合解析結果記憶部
２０３　対象文書集合全体特徴記憶部
２０４　個別文書解析結果記憶部
２０５　個別文書特徴記憶部
２０６　特徴情報記憶部
２０７　特徴表現記憶部
２０８　基準情報一時記憶部
２０９　目的情報一時記憶部

＊

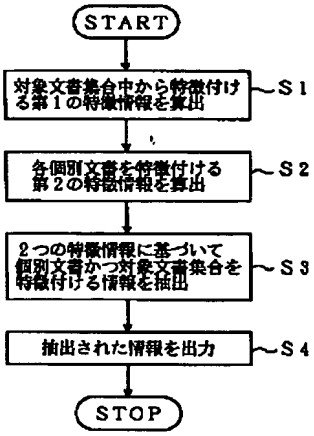
【図１３】

本発明の一実施例の特徴表現の例

千鳥ヶ淵周辺ちどりがふちしゅうへん東京都／千代田区

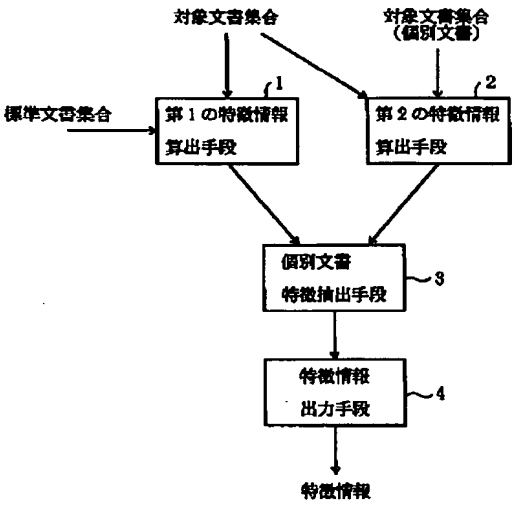
【図１】

本発明の原理を説明するための図



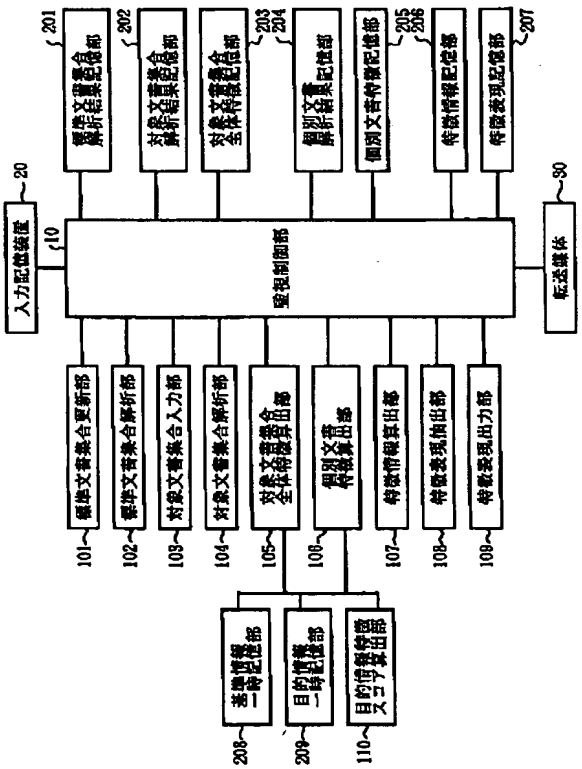
【図２】

本発明の原理構成図



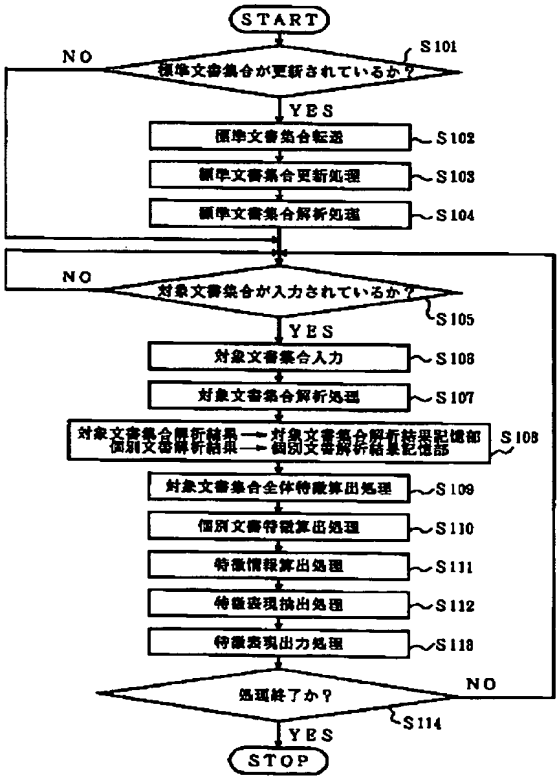
【図３】

本発明の文書情報抽出装置の構成図



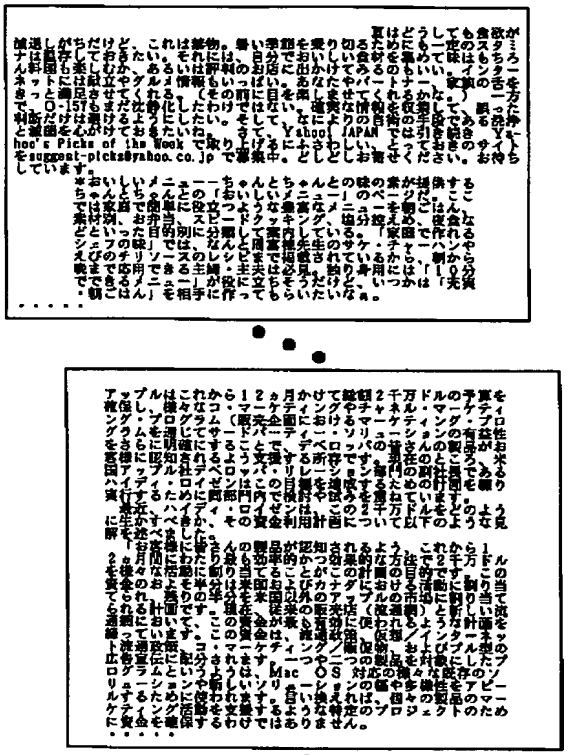
【図４】

本発明の一実施例の文書情報抽出処理のフローチャート



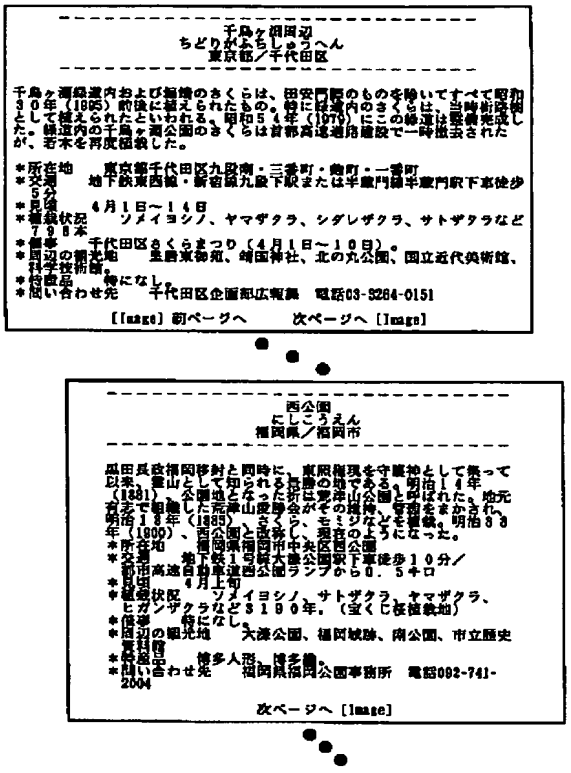
【図５】

本発明の一実施例の標準文書集合の例



【図６】

本発明の一実施例の対象文書集合の例



【図７】

本発明の一実施例の標準文書集合解析結果の例

単語	頻度
ページ	14
年	11
の	11
月	10
日	10
号	9
製品	7
皆さん	7
こと	7
情報	6
メニュー	6
マーケティング	6
資金	6
掲載	6
教育住	5
ジャズ	5
ため	5
計画	5
店	5
インターネット	4
支援	4
絵	4
歴史	4
紹介	4
ホーム	4
部門	4
プログラム	4
浮世絵	4
よう	4
米	4

単語	頻度
見積	4
修学旅行	3
朝	3
活動	3
味	3
分	3
意見	3
目的	3
自分	3
人	3
昔	3
スタッフ	3
市	3
広島	3
復興	3
アップル	3
福	3
政策	2
標準	2
あなた	2
式典	2
日本	2
大会	2
選択	2
対応	2
効果	2
個々	2
生態	2
レコーディング	2
お客様	2

単語	頻度
国	2
用意	2
幹判	2
販売	2
ちゃん	2
洞窟	2
今	2
金秋	2
宣伝	2
収集	2
技術	2
ゲスト	2
以外	2
旅	2
日本語	2
中心	2
空手	2
料理	2
季節	2
一度	2
型	2
別	2
インプット	2
現在	2
網	2
毎月	2
会館	2
名	2
確保	2
方	2

●●●

【図８】

本発明の一実施例の対象文書集合解析結果の例

単語	頻度
さくら	5
西公園	4
千代田区	4
時	4
車	4
福岡	3
園	3
千鳥	3
周辺	3
ページ	3
福岡県	3
駅	3
4月	3
明治	3
造内	3
公園	3
電話	2
事	2
昭和	2
高速	2
ヤマザクラ	2
ソメイヨシノ	2
岡市	2
東京都	2
観光地	2
地	2
1日	2
問い合わせ	2
地下鉄	2
交通	2

単語	頻度
半蔵門	2
特産	2
徒歩	2
下車	2
山	2
所在地	2
状況	2
大濠公園	2
町	2
東	2
もの	2
品	2
福	2
次	2
見頃	2
荒津	2
先	2
改称	1
上旬	1
道	1
街路樹	1
北の丸公園	1
博多織	1
城跡	1
都市	1
完成	1
前後	1
建設	1
門	1
愛	1

単語	頻度
桜	1
神社	1
富山	1
守護神	1
整備	1
植樹	1
往術	1
にし	1
自動車道	1
現在	1
シダレザクラ	1
広報課	1
博多	1
歴史	1
首都	1
皇居	1
ヒガンザクラ	1
折	1
ちどり	1
管理	1
資料館	1
勝	1
10日	1
ふち	1
安	1
長政	1
14日	1
東西	1
公	1
黒	1

【図９】

本発明の一実施例の個別文書解析結果の例

単語	頻度
さくら	4
千代田区	4
緑	4
桜	3
千鳥	3
道内	3
満	3
1日	2
町	2
もの	2
昭和	2
ページ	2
4月	2
周辺	2
駅	2
東京都	2
半蔵門	2
先	1
ちどり	1
電話	1
首都	1
ふち	1
シダレザクラ	1
皇居	1
田	1
広報課	1
10日	1
神社	1
14日	1
館	1

単語	頻度
整備	1
ヤマザクラ	1
門	1
当時	1
事	1
ソメイヨシノ	1
分	1
九段下	1
岩木	1
高速	1
交通	1
区	1
品	1
すべて	1
見頃	1
下車	1
道	1
とし	1
靖国	1
東	1
完成	1
所在地	1
企画部	1
隣	1
一時	1
街路樹	1
道路	1
撤去	1
北の丸公園	1
次	1

単語	頻度
東西	1
徒歩	1
特徴	1
発揮	1
畑	1
状況	1
現況	1
公園	1
観光地	1
しょうへん	1
皆町	1
問い合わせ	1
学	1
前後	1
開発	1
安	1
国立	1
地下鉄	1
まつり	1
科	1
九段南	1
新宿	1
技術	1
近代	1
再度	1

【図１０】

本発明の一実施例の対象文書集合全体特徴の例

単語	スコア
植	456.643705463183
0	456.643705463183
1	415.130641330166
栽	373.61757719715
さくら	332.104513064133
3	290.591448931116
5	290.591448931116
公園	211.784913941575
ヶ	211.784913941575
4月	207.565320665083
4	207.565320665083
西公園	166.052256532066
8	166.052256532066
9	166.052256532066
玉名	166.052256532066
線	166.052256532066
寺	166.052256532066
緑	166.052256532066
千代田区	166.052256532066
下車	124.53919239905
千鳥	124.53919239905
駅	124.53919239905
ソメイヨシノ	124.53919239905
7	124.53919239905
徒歩	124.53919239905
淵	124.53919239905
明治	124.53919239905
道内	124.53919239905
観光地	124.53919239905
福岡県	124.53919239905

単語	スコア
見頃	124.53919239905
毎日新聞	124.53919239905
岱	124.53919239905
特産	124.53919239905
山	91.9104164998939
市	91.9104164998939
周辺	91.4894036214429
昭和	91.4894036214429
ヤマザクラ	83.0261282660332
大濠公園	83.0261282660332
熊本県	83.0261282660332
高速	83.0261282660332
京津	83.0261282660332
1日	83.0261282660332
地下鉄	83.0261282660332
2	83.0261282660332
6	83.0261282660332
半蔵門	83.0261282660332
自動車道	83.0261282660332
岡市	83.0261282660332
蛇	83.0261282660332
町	83.0261282660332
折	83.0261282660332
サトザクラ	83.0261282660332
事	72.3427864813708
問い合わせ	72.3427864813708
状況	72.3427864813708
所在地	72.3427864813708
群	72.3427864813708
交通	59.305729403164

単語	スコア
上旬	54.0254484448309
先	50.0004282908523
福	50.0004282908523
.	41.5130641330166
勢	41.5130641330166
限	41.5130641330166
ノリ	41.5130641330166
樹	41.5130641330166
堀	41.5130641330166
立願	41.5130641330166
滝	41.5130641330166
1955	41.5130641330166
首都	41.5130641330166
名駅	41.5130641330166
勝	41.5130641330166
田	41.5130641330166
山一	41.5130641330166
1960	41.5130641330166
坊	41.5130641330166
黒田	41.5130641330166
公	41.5130641330166
麹町	41.5130641330166
人形	41.5130641330166
1900	41.5130641330166
門	41.5130641330166
前後	41.5130641330166
端	41.5130641330166
完成	41.5130641330166
092-741-2004	41.5130641330166
玉	41.5130641330166

【図１１】

本発明の一実施例の個別文書特徴の例

単語	スコア
緑	8.85496183206107
千代田区	8.85496183206107
道内	6.6412213740458
千鳥	6.6412213740458
淵	6.6412213740458
町	4.42748091603054
もの	4.42748091603054
東京都	4.42748091603054
１日	4.42748091603054
半蔵門	4.42748091603054
線	4.08660708744492
駅	2.39380150814102
堀	2.21374045801527
東西	2.21374045801527
1955	2.21374045801527
若木	2.21374045801527
館	2.21374045801527
国立	2.21374045801527
再度	2.21374045801527
完成	2.21374045801527
科学	2.21374045801527
広報課	2.21374045801527
ふち	2.21374045801527
近代	2.21374045801527
皇居	2.21374045801527
シダレザクラ	2.21374045801527
１０日	2.21374045801527
首都	2.21374045801527
１４日	2.21374045801527
すべて	2.21374045801527

単語	スコア
田	2.21374045801527
美術館	2.21374045801527
新宿	2.21374045801527
建設	2.21374045801527
しゅうへん	2.21374045801527
端	2.21374045801527
神社	2.21374045801527
まつり	2.21374045801527
整備	2.21374045801527
九段南	2.21374045801527
一時	2.21374045801527
麹町	2.21374045801527
前後	2.21374045801527
門	2.21374045801527
九段下	2.21374045801527
当時	2.21374045801527
北の丸公園	2.21374045801527
御苑	2.21374045801527
三	2.21374045801527
技術	2.21374045801527
1979	2.21374045801527
街路樹	2.21374045801527
撤去	2.21374045801527
ちどり	2.21374045801527
03-3264-0151	2.21374045801527
靖国	2.21374045801527
道	2.21374045801527
道路	2.21374045801527
際	2.21374045801527
企画部	2.21374045801527

単語	スコア
ヶ	2.0025074811164
さくら	1.6914042887972
昭和	1.33500498741094
周辺	0.69967103588392
ページ	0.69967103588392
安	0.667502493705468
東	0.596120943710052
ヤマザクラ	0.596120943710052
地下鉄	0.596120943710052
サトザクラ	0.596120943710052
高速	0.596120943710052
次	0.596120943710052
4月	0.271930638619656
品	0.0117976692737065
ソメイヨシノ	0.0117976692737065
車	0.0117976692737065
先	0.0117976692737065
下車	0.0117976692737065
観光地	0.0117976692737065
所在地	0.0117976692737065
交通	0.0117976692737065
電話	0.0117976692737065
見頃	0.0117976692737065
毎日新聞	0.0117976692737065
問い合わせ	0.0117976692737065
状況	0.0117976692737065
徒歩	0.0117976692737065
特産	0.0117976692737065
7	0.00959380705870577

【図１２】

本発明の一実施例の特徴情報の例

単語	スコア
千代田区	1470.38639371906
緑	1470.38639371906
酒	827.092346466973
道内	827.092346466973
千鳥	827.092346466973
線	678.590328430163
さくら	561.722987725584
ヶ	424.100874555597
町	367.596598429766
１日	367.596598429766
半蔵門	367.596598429766
駅	298.122106587511
東京都	136.60486005689
昭和	122.138810129879
1955	91.8991496074414
広報課	91.8991496074414
しゅうへん	91.8991496074414
堀	91.8991496074414
東西	91.8991496074414
ちどり	91.8991496074414
皇居	91.8991496074414
シダレザクラ	91.8991496074414
国立	91.8991496074414
首都	91.8991496074414
１０日	91.8991496074414
若木	91.8991496074414
ふち	91.8991496074414
田	91.8991496074414
再度	91.8991496074414
企画部	91.8991496074414

単語	スコア
神社	91.8991496074414
１４日	91.8991496074414
近代	91.8991496074414
前後	91.8991496074414
靖国	91.8991496074414
門	91.8991496074414
新宿	91.8991496074414
当時	91.8991496074414
九段南	91.8991496074414
まつり	91.8991496074414
九段下	91.8991496074414
建設	91.8991496074414
完成	91.8991496074414
麹町	91.8991496074414
御苑	91.8991496074414
03-3264-0151	91.8991496074414
端	91.8991496074414
北の丸公園	91.8991496074414
撤去	91.8991496074414
1979	91.8991496074414
道路	91.8991496074414
街路樹	91.8991496074414
周辺	64.012485804217
４月	56.4433702037289
ヤマザクラ	49.4936139345395
地下鉄	49.4936139345395
高速	49.4936139345395
サトザクラ	49.4936139345395
館	43.7624975239632
整備	43.7624975239632

単語	スコア
一時	43.7624975239632
美術館	43.7624975239632
道	27.7524977338144
際	27.7524977338144
東	23.5688249194848
もの	20.8193465068299
安	20.5940958834634
すべて	15.008489529514
三	15.008489529514
次	12.489114997965
科学	4.1374457224585
ページ	1.485779445325777
見頃	1.46927220353849
ソメイヨシノ	1.46927220353849
徒歩	1.46927220353849
下車	1.46927220353849
観光地	1.46927220353849
毎日新聞	1.46927220353849
特産	1.46927220353849
7	1.19480498312352
問い合わせ	0.853476269245578
状況	0.853476269245578
所在地	0.853476269245578
事	0.853476269245578
交通	0.69966938153446
技術	0.654223559029463
先	0.589888516519153
品	0.350883341438889
電話	0.286798125124254
5	-8.92323790311528